THE PROCEEDINGS OF
THE PHYSICAL SOCIETY

VOL. 59, PART 2 1 March 1947 No. 332

A NOTE ON THE INTERMITTENCY EFFECT

By R. WEIL,
South-West Essex Technical College

Communicated by Dr. H. Lowery; *MS. received* 10 *July* 1946

ABSTRACT. The nature of the intermittency effect is described and a new criterion for its detection mentioned. This is applied to the manifestation of the effect when an alternating current light source is used. The experimental arrangement employed in the present investigation is given, and the results of the latter are briefly discussed.

§1. INTRODUCTION

NUMEROUS workers have found that photographic materials are, in general, unreliable in integrating an intermittent exposure. Thus the density produced on a plate by ten one-tenth-of-a-second exposures will differ from that due to a single exposure of the same intensity lasting one second. This fact may be found when both the intermittent and the continuous exposures are made on one plate: it is due to an objective phenomenon resulting from the way in which light reaches the individual grains of the emulsion, and is known as the *Intermittency Effect.*

In connection with intermittent exposures it has also been found that the Bunsen-Roscoe reciprocity law breaks down. According to the latter, $I \times t =$ constant, where I is the intensity of the light and t is the time of exposure. This law applies only to the straight-line portion of the curve connecting the density d and the logarithm of the exposure. Its breakdown implies that the ratio of two densities does not equal the ratio of the logarithms of the corresponding exposures (exposure $E = I \times t$).

The intermittency of illumination is generally produced by means of a rotating sector-wheel. Talbot was the first to suggest that if such a wheel were rotated at a suitably high velocity the eye would obtain the impression that any light behind the wheel was of intensity I', whereas its true intensity was I. The ratio I'/I would be given by

$$I'/I = \theta/2\pi, \qquad \ldots\ldots(1)$$

where θ is the angle of the sector.

In a series of brilliant researches, Webb demonstrated the true nature of the intermittency effect. He found that it was closely connected with the failure of the Bunsen-Roscoe reciprocity law. It should be pointed out that, because of these effects, it has become necessary to introduce two scales in photographic measurements: the intensity-scale in which the intensity is variable and the time-factor constant, and *vice versa* for the time-scale (figure 1). Webb found that the intermittent time-scale curve, which lies generally between the two others, coincides with the former if the sector-wheel is run above a critical speed. (Webb and Silberstein (1934) have later shown that such a coincidence is not possible theoretically but may be considered to exist for all practical purposes.) If light is propagated in discrete quanta, it is evident that any exposure is really intermittent. On the basis of this conception, Webb showed that the critical speed mentioned above was such that, on the average, not more than one quantum of light per flash falls on one grain of the photographic emulsion. The critical speed was given by the relation

$$f_c = g \cdot I_{av}, \qquad \ldots\ldots(2)$$

where g is the effective reception area of the grain and I_{av} the average intensity.

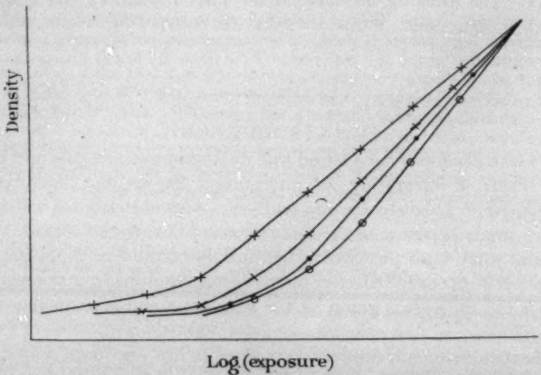


Figure 1. + Time scale. × Low frequency of flash. ● High frequency of flash. ○ Intensity scale.

This equation states that the critical frequency is proportional to the average intensity. It is plain that the probability of one quantum being transmitted by the sector-wheel is greater for a given frequency of flash if the sector-aperture is larger. It is therefore to be expected that if two different sectors are run at the same speed (cf. Lochte-Holtgreven and Maecker (1937)), the intermittency effect will or will not be apparent according as the speed is below or above the critical speed for the larger sector. Since rotating sector-wheels are used in measurements of reflectivity and absorptivity, and, therefore, must have variable apertures, it is vital that the above point should be borne in mind.

§ 2. EXPERIMENTAL ARRANGEMENT

In order to investigate the possibility of using A.C. sources for photometry by photographic methods, the author designed a special sector-wheel (figure 2). As has been mentioned before, rotating sectors are used in measurements of absorption and reflectivity: the beam which is not absorbed or reflected is passed through the wheel and the sector of the latter adjusted until a match is obtained between its intensity and that of the other beam. When the speed of the wheel is very high it is necessary to ensure that it should be as symmetrical as possible to prevent it from breaking owing to centrifugal forces. That is why two sectors facing each other were cut into the wheel.

It has also been shown that if both beams are interrupted at the same rate, the intermittency effect will be eliminated provided that the frequency of flash corresponds to, or is higher than, the critical frequency for the larger aperture. Lochte-Holtgreven and Maecker, who used a rotating sector-wheel in their work on the temperature of a freely burning carbon arc, employed a dummy sector in the path of the standard beam to eliminate the effect. In the present arrangement only one wheel is used, and both beams are passed through it: the beam to be reflected or absorbed through the large angle, which is fixed (at 80°), and the other through the variable sector. The latter has a radius only one half that of the wheel, and therefore cannot interfere with the passage of the beam through the large sector. For a given intensity there will be a marked intermittency effect at low frequencies of flash. But for every flash the current will alternate many times and, on the average, the stroboscopic effect will be negligible. As the speed of rotation is increased, the intermittency effect proper will be eliminated, but there will be fewer current-cycles per flash and a secondary effect would become apparent. In fact, it was seen that when the sector-speed was 32 r.p.s.—a speed at which no individual flashes can be perceived—a distinct flicker was observed when the A.C. source was viewed: sunlight, however, reflected from a mirror, and transmitted

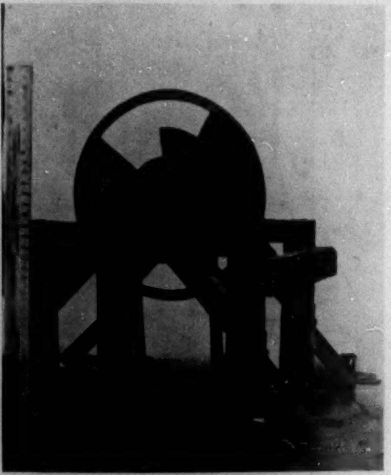


Figure 2.

through the sector-wheel running at the same speed, appeared to be perfectly steady. This secondary effect would be expected to affect the photographic plate in exactly the same way as the ordinary effect in connection with D.C. sources of light.

It is seen from figure 1 that the intermittency effect gives rise to a density difference for equal exposures owing to the existence of different curves for the two beams if only one is interrupted . Thus, if the difference in density is plotted against the corresponding difference in the logarithms of the exposures for one standard beam, curves will be obtained which do not pass through the origin if the intermittency effect is present. Conversely, if they pass through the origin its elimination may be assumed. In particular, if the straight-line portions of both curves are used—a condition which can be fulfilled only with plates having a small γ—straight lines will result whose tangent is equal to γ, but which will pass through the origin only if the intermittency effect is eliminated. This will be the case when the exposures due to both beams lie on the same curve. Although the theory does not apply to heavily over-exposed plates, it is evident from figure 1 that in the case of too long or too short exposures, all the curves coincide and no conclusive results can be obtained. This fact is only mentioned because it is not essential that our measurements be confined to the straight-line portion of the curve, although this is desirable. Indeed, as can be seen from figure 3, with very contrasty plates it is practically impossible to work on the straight-line portion only.

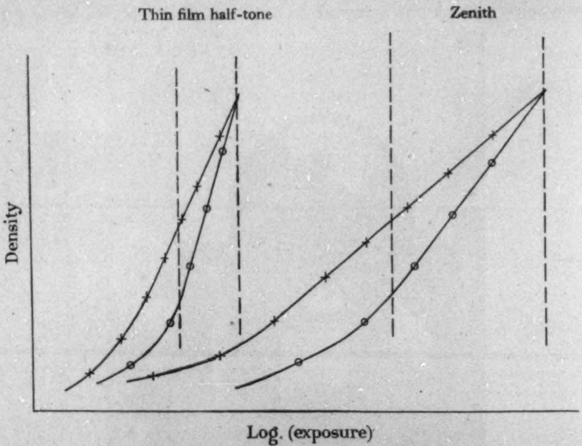


Figure 3. Range of straight line portion.

Since the exposure times are proportional to the two sector apertures, the densities produced by the two parts of the wheel will be given by the expressions :

$$d_1 = k \cdot \log I\alpha_1 + b, \qquad \ldots\ldots(3)$$

$$d_2 = k \cdot \log I\alpha_2 + b, \qquad \ldots\ldots(4)$$

where α_1 and α_2 represent the angular apertures of the sectors and b is a factor representing any accidental fogging. Since the two blackened areas on the plate are near to each other, b may be assumed to be the same for each of them. Subtracting the second equation from the first,

$$d_1 - d_2 = k \, . \log \frac{\alpha_1}{\alpha_2}. \qquad \ldots\ldots(5)$$

This expression refers only to the straight-line portion of the curve. But if an arbitrary point is taken on the intensity-scale curve in figure 1, it will be obvious that, although no straight-line graphs need necessarily be obtained, the curve will pass through the origin if the intermittency effect is absent.

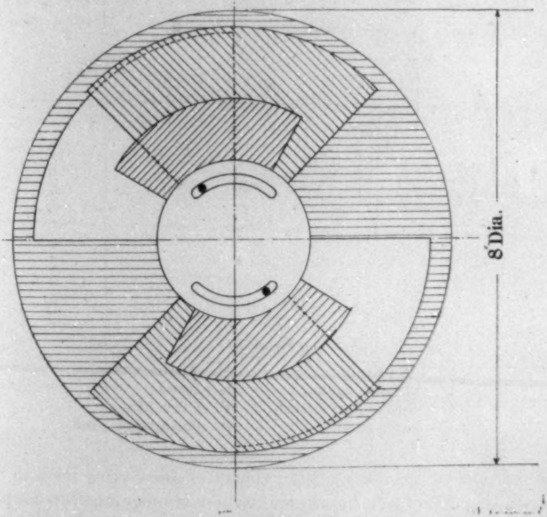


Figure 4.

The sector-wheel consisted of a brass plate 8″ in diameter and $\frac{1}{16}$″ thick, with two sectors cut into it diagonally opposite each other (figure 4). A large vane allowed the sector-apertures to be reduced from 90° to 0°: during the subsequent investigation the apertures were kept at 80°, as this facilitated the procedure. Another vane, similar to the previous but with only half the diameter, permitted the further cutting down of the central parts of the apertures. The whole arrangement was painted black. A ground-glass screen illuminated with the diffused light of a filament lamp was placed immediately in front of the wheel (figure 5). A lens focused the screen on the photographic plate, and a green filter (Ilford 404) was interposed. It was desirable to use monochromatic light since, according to Webb, the critical frequency varied with the wave-length of light. The sector-wheel was driven by an electric motor whose speed could be varied. The two turned at the same speed, which was measured on the shaft of the motor by means of a tachometer.

Before the actual experiment was embarked upon, the characteristic curves of the two plates in use were determined so that extremes of exposure could be avoided.

As mentioned above, the aperture α_1 was kept at 80° throughout the experiment. On any one plate, for which α_2 was kept constant, eight photographs were taken for frequencies of flash ranging from 4 to 32 per second. Since only the difference in densities is considered, it would appear to be in order to combine the results for different samples of one kind of plate. Various objections might be raised in connection with this procedure, but it is difficult to justify an alternative. The object of the investigation was to find variations which depended on the frequency of flash. Evidently there was a better chance of discovering them in the above manner than if different sector-apertures had been photographed on one plate and a different speed of rotation chosen for every plate. A test for the reliability of the procedure was afforded by the regularity of the densities of the patches blackened by a continuous exposure, as will be seen later. In order to demonstrate the justification of the use of a single wheel for the two beams, a series of photographs was taken in which only one beam was interrupted, the other subjecting the plate to a continuous exposure. The ratio α_1/α_2 was taken to be $2\pi/\theta$. Two plates were used: Zenith (H & D 700) and Ilford Thin Film Half-Tone. Their relative slopes are indicated in figure 3, and it is evident that the range for work along the straight-line portion is very small in the latter case. The times of exposure varied from 40 to 100 seconds, but were kept constant for any particular series of exposures. The usual development technique was used. In order to reduce any error due to the use of different samples of one kind of plate, the times of development were chosen so as to obtain a maximum γ, in fact γ_∞. That this has been achieved has already been stressed above.

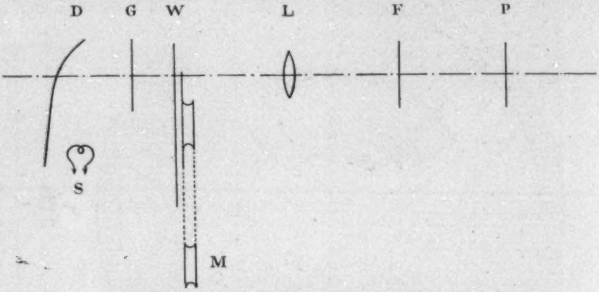


Figure 5.

S.	Light source.	M.	Electric motor.
D.	Diffusing screen.	L.	Lens.
G.	Ground-glass screen.	F.	Green filter.
W	Sector-wheel.	P.	Photographic plate.

The plates were subsequently examined with a microphotometer and graphs plotted as suggested above.

§ 3. RESULTS

In the cases when only one part of the exposure was intermittent (and the other continuous) irregular results were obtained for the Zenith plates; those for the Thin Film are shown in figure 6. It is clear that a density difference exists for the lower frequencies of flash and that, at higher frequencies, the curves become irregular owing to secondary effects which are connected with the alternating source of light.

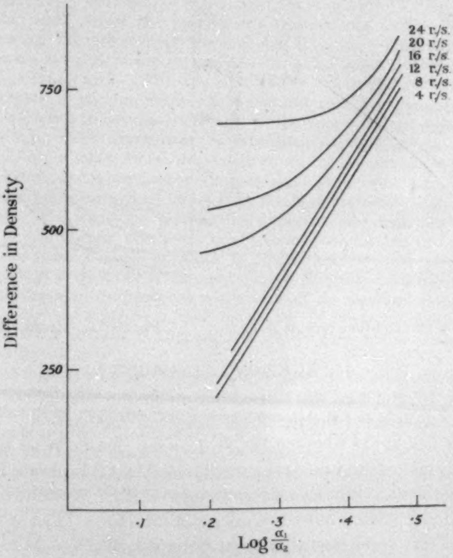


Figure 6. Thin film half-tone (single sector).

But with both beams interrupted at the same rate, there is an intermittency effect at the lower frequencies only, and this disappears as the speed of rotation is raised (figures 7 *a*, 7 *b*). This is in agreement with expectations. The fact that the critical frequencies for both kinds of plate lie in the neighbourhood of 12 r.p.s. is probably accidental as a different lamp was used for each kind. At frequencies of 28 flashes per second and more, the intermittency effect reappeared. The curves representing 28 and 32 r.p.s. are not indicated in the figures as they would tend to obscure rather than clarify the point in question. Let it be said, however, that they do not pass through the origin but approach the positions of the curves for less than 16 flashes per second. The stroboscopic effect, so obvious to the eye, is also affecting the photographic plate.

§ 4. CONCLUSION

Webb's work has shown that the intermittency effect can be eliminated when D.C. sources of light are used if the frequency of flash is above a certain critical value. From the above work it is seen that the same applies to A.C. sources on these conditions: (*a*) the standard beam and the beam to be reflected or absorbed must be interrupted at the same rate; (*b*) the frequency of flash must be greater than a certain critical frequency; (*c*) the frequency due to the stroboscopic effect, i.e. the combination of the frequencies of the A.C. source and the sector wheel, must be above the same critical frequency. This latter statement was substantiated by visual observations.

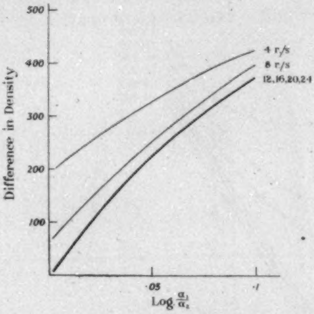


Figure 7 *a*. Thin film half-tone (double sector).

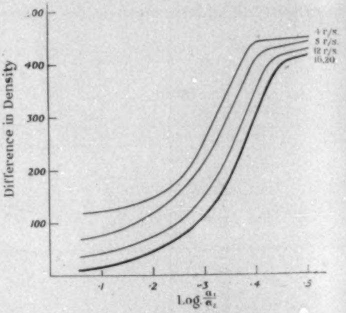


Figure 7 *b*. Zenith (double sector).

§ 5. ACKNOWLEDGMENTS

The above investigation was carried out under the auspices of the Pyrometry Sub-Committee, British Iron and Steel Research Association.

The writer would like to thank Dr. H. Lowery, South-West Essex Technical College, for his interest and encouragement, and Mr. G. Lothian, M.A., for many valuable discussions. Sir Lawrence Bragg, F.R.S., Cavendish Laboratory, Cambridge, kindly made available a microphotometer. The work would have been rendered very cumbersome without the good offices of Miss M. MacIntyre and Mr. W. Bennett.

REFERENCES

Forsythe, 1940. *Measurement of Radiant Energy* (McGraw Hill), Chapter VIII *.
Lochte-Holtgreven and Maecker, 1937. *Z. Phys.*, **105**, 1.
Silberstein and Webb, 1934. *Phil. Mag.*, **18**, 1.

* This chapter contains a very extensive and up-to-date bibliography on the intermittency effect.

DELAYED FRACTURE IN GLASS

By C. GURNEY,
Royal Aircraft Establishment, Farnborough

MS. received 6 *August* 1946

ABSTRACT. An important delayed-fracture effect is found for mineral glass. A likely cause might be the gradual spread of Griffith cracks, but a theorem due to Griffith states that cracks do not spread below a certain average stress, and that at this stress they spread catastrophically. In the present paper, Griffith's theorem is re-examined and it is shown that in materials having atomic constitution, fracture does not occur catastrophically at the Griffith stress ; the Griffith stress is the least at which a crack may start to spread by a process of splitting, and the rate of spread is controlled by the rate at which thermal motions overcome energy barriers. Catastrophic fracture does not occur until the stress at the end of the crack equals the maximum a material can withstand. An improved method of estimating this stress from thermal data is given. The best estimate is that the maximum stress is equal to the intensity of the given stress system which makes the latent heat of evaporation zero. The rate of spread of cracks by Griffith's process is considered to be too slow to account for much of the delayed-fracture effect in glass. Other processes are considered under the headings of approach to homogeneous and heterogeneous equilibrium. The attainment of homogeneous equilibrium under stress in materials in equilibrium when stress-free involves, in the absence of phase changes, but a small entropy change, and is not likely to cause an appreciable time effect. Glass, however, is not in thermal equilibrium when stress-free, the high temperature phase persisting at temperatures below the transition point. On account of the high internal viscosity of glass, approach to equilibrium effectively ceases soon after manufacture. Stress reduces the internal viscosity and enables approach to equilibrium to continue. Because of the concentration of stress at the ends of the cracks, the approach to equilibrium made possible by stress is much faster in the material at the ends of the cracks than elsewhere, and as attainment of equilibrium involves volume shrinkage, the stress at the end of the crack is increased and the crack spreads. This effect would be expected to cause an appreciable delayed fracture effect.

Two effects are considered under the heading of approach to heterogeneous equilibrium. The first is evaporation of the material at the end of the crack. An estimate of delayed fracture due to this cause suggests that it is unimportant. A much more important cause of delayed fracture is atmospheric attack of the glass. Due to concentration of strain energy, the material at the end of the crack has a much higher free energy than normal unstressed glass, and is therefore much more chemically active. Atmospheric attack will result in the formation of a complex of glass and atmospheric constituents. The crack will extend continually if the strength of this complex, during or after its formation, is less than the load imposed on it.

Changes due to stress in the stable phase at room temperature are considered, but these are not likely to be important for glass, as in this material the high-temperature phase persists at room temperature. Phase changes under stress may have important effects on the behaviour of other materials. Included in the paper are formulae for the stress coefficients of vapour pressure, entropy and phase-transition temperature, of material subjected to a generalized stress system.

§ 1. INTRODUCTION

A VERY remarkable delayed-fracture effect is found in mineral glass, the time to fracture increasing with decreasing load, the earliest investigation being due to Grenet (1899). More recently Preston (1942),* quoting results obtained by T. C. Baker, found that the time for annealed soda lime glass to fracture in bending increased from 10^{-2} sec. to 10^{4} sec. when the load was reduced in the ratio 3 : 1. Somewhat similar results were obtained by Holland and Turner (1940). According to Griffith (1920), whose theory has been accepted by many subsequent investigators, glass contains submicroscopic cracks, and its comparative weakness is due to concentration of stress at the ends of these cracks. A likely explanation of the delayed fracture of glass would be the gradual growth of these cracks in material under load. Griffith, however, considered the conditions for crack growth, and decided that, below a critical load, crack growth could not occur, and that at this load the crack spread catastrophically. The critical load was that at which the spreading of the crack would result in a reduction of the free energy of the system.

---

* Since this paper was written a fuller account of the work of Preston and his collaborators has become available. It is given in papers by Preston, Baker and Glathart in *J. Appl. Phys.*, **17**, 162 (1946).

Recent attempts at the explanation of the delayed fracture of glass have been consistent with Griffith's theorem. Orowan (1944) has suggested that the main effect may be caused by atmospheric attack of the surface of the glass. He suggests that the surface tension of contaminated glass may be about one-tenth that of a freshly formed glass surface. Griffith's energy theorem gives the stress to fracture to be proportional to the square root of the surface tension, so that contamination of the surface reduces the stress at which the crack spreads in the ratio of about 3 : 1. When the load applied to glass is sufficient to break it quickly there may be no time for contamination of the rapidly formed surfaces of the spreading crack. In this case, a stress corresponding to the higher surface tension of the glass is obtained. If the load is not sufficient to break the glass quickly, but exceeds that necessary to cause a crack to spread if the surface tension has its lower value, the crack will gradually spread. The rate of spreading will depend on the time for the contamination of the freshly formed surface to occur to an extent sufficient to reduce the surface tension to that necessary for the crack to spread.

Orowan's explanation of the delayed-fracture effect implies that the strength of glass as measured in an ordinary quick-loading test will be reduced by preloading the glass at a load sufficient ultimately to break it. Murgatroyd and Sykes (1945), who experimented with preloaded test pieces, found no significant reduction in strength, and Murgatroyd has proposed an alternative explanation of the delayed-fracture effect. Murgatroyd (1944) has suggested that the Griffith cracks are filled with a viscous constituent of the glass which gradually relaxes under load. At the commencement of loading there is no concentration of stress at the ends of the cracks, but as the viscous material relaxes, concentration of stress at the ends of the cracks increases until fracture occurs. The present author in collaboration with Pearson (1946) has made experiments on the delayed fracture of round glass rods under four-point bending. Three series of tests were made. In the first the test pieces were not rotated and in the second and third the test pieces were rotated at 14 and 10,000 r.p.m. respectively, the direction of loading being stationary. The curves relating stress and time to fracture for the three conditions did not differ significantly. This result is at variance with Murgatroyd's theory. Under cyclic stress, the viscous constituent of Murgatroyd's model would not be expected to relax continuously, and delayed fracture under cyclic stress would not be expected.

It seemed to the present author that Orowan's and Murgatroyd's explanations were restricted by their acceptance of the Griffith load as being that at which cracks spread catastrophically, and below which cracks did not spread at all. In this paper Griffith's theory is reviewed, and it is concluded that it is consistent with the gradual spread of cracks even in the absence of atmospheric attack. Catastrophic fracture does not occur until the stress at the end of the crack reaches the maximum the material can withstand. § 2 of this paper therefore considers the estimation of this stress. In § 3, crack spreading by splitting in the manner envisaged by Griffith is treated. In §§ 4 and 5 other processes leading to delayed fracture are suggested. These result from the stressed material approaching thermal equilibrium. It has been found convenient to distinguish effects related to homogeneous equilibrium from those related to heterogeneous equilibrium. Homogeneous equilibrium is concerned with the arrangement of atoms in a given phase. As the strength of material depends on the atomic arrangement, gradual rearrangement of atoms will result in a gradual change of strength. Heterogeneous equilibrium is concerned with the partition of material amongst the various phases, as for example the number of atoms in the vapour phase in a material in equilibrium with its vapour; evaporation of material at the ends of cracks would lead to cracks spreading and delayed fracture. In § 6, phase changes caused by stress are considered.

Before proceeding, the application of thermodynamics to the minute quantity of highly stressed material at the bottom of the cracks needs consideration. The laws of thermodynamics are statistical laws, and strictly apply only to assemblies containing large numbers of elementary units. As the number of units is reduced, the variations of the results of individual experiments from the mean result increases, the standard error varying inversely with the square root of the number of units, but no change in quality of the laws appears until very few repetitions of experiments involving very few atoms are considered. In the latter case the laws cease to hold; for example, in a crystalline material, the jump of an atom from a lattice point to an interstitial position can occur although it causes an increase in free energy. In the case of cracks spreading, however, movements of quite large numbers of atoms are involved. Griffith cracks in glass are estimated to be of the order of 10^{-3} cm. long. Stress is proportional to the square root of crack length, so that to account for delayed fracture at 70% of the quick-loading strength, the cracks must double their length. If only the single string of atoms along the longest line contour of the crack is affected, the number of atoms involved in doubling the length of the crack will be of the order of 10^{10}, so that statistical laws should apply without very great variation in results.

Another subject which needs a brief discussion is the application of thermodynamics to stress systems other than hydrostatic pressure. Formulae for the stress gradient of vapour pressure of material under a generalized stress system will be given later. In general, the vapour pressure differs on each of the three pairs of parallel faces which bound a cube of material, and the system therefore cannot be in equilibrium. If, however, the time for any effective change to be caused by evaporation is long compared with the time for some other change, e.g. chemical combination, to occur, the condition can be regarded as one of metastability, and thermodynamic formulae can be applied. It may be worth mentioning here that the free energy of any thermodynamic system involving other than hydrostatic pressure may be reduced by the material spontaneously separating or shearing along internal surfaces; because the work done by the external forces exceeds the surface energy of the fracture surfaces. Usually, however, the time for spontaneous fracture is so long compared with the time for fracture by crack spreading, that spontaneous fracture is not a factor of any practical importance.

§2. THE STRENGTH OF FLAWLESS MATERIAL

The strength of flawless materials is usually estimated from thermal data such as latent heat, surface tension and triple point or from a theory of atomic binding. A summary of this work is given by Houwink (1937). The estimate of strength given here does not differ greatly in principle from previous estimates, but it gives them more precise expression.

A thermodynamic system under uniform pressure becomes unstable when the partial derivative of the pressure with respect to the volume becomes positive. By similar reasoning, material under a generalized stress system will become unstable when the partial derivative of the load in one direction with respect to the movement in that direction becomes negative, load and movement being of the same sign when in the same direction. Instability in a liquid under hydrostatic pressure results in the sudden formation of two phases. In a solid under generalized stress, instability with respect to a tensile stress results in fracture. In terms of atomic constitution, fracture would be expected to occur virtually immediately on loading, if the average thermal motion was just sufficient to enable atoms associated with the average strain energy to reach the top of the average energy barrier. If more than the average thermal motion is necessary, fracture may still occur, but it will not be immediate. Individual atoms will occasionally acquire sufficient kinetic energy to enable them to migrate, leaving a hole which by the same process will gradually develop into a crack and lead to fracture.

Physical knowledge of glass is not at present adequate for the above criterion to be used to estimate the maximum strength of flawless material. It is however possible to estimate the stress at which the average thermal motion is sufficient to vaporize the material, for this is the stress at which the latent heat of isothermal evaporation is zero. This stress may be expected to exceed the true breaking stress by a factor of two or more, but it seems the best estimate of theoretical strength at present possible.

The calculation of the stress at which the latent heat of evaporation is zero may be made by formal thermodynamic methods. It is first necessary to calculate the change of vapour pressure with stress. A three-dimensional stress system can be denoted by X_{ij}, where i and j take values $1, 2, 3$; the corresponding strain system can be denoted by e_{ij}. The differential of strain energy per unit volume is then

$$\Sigma X_{ij} de_{ij}. \qquad \ldots\ldots(1)$$

This expression is the change of internal energy due to external work done on the material. It replaces the more familiar pdv, which applies to the hydrostatic pressure case. The calculation proceeds by stating that at equilibrium a small change in the system must be reversible, and that in a reversible isothermal process, change in the Helmholtz free energy is equal to the external work done on the system. It is found that the vapour pressure varies with the direction of the surface from which evaporation is considered to take place. If any particular stress is denoted by X_{kl}, the partial change in vapour pressure (p_{mm}) of the surface whose normal is mm for $kl \neq mm$ is given by

$$\frac{\partial p_{mm}}{\partial X_{kl}} = \frac{\Sigma_{ij} X_{ij} \dfrac{\partial e_{ij}}{\partial X_{kl}} - X_{mm} \dfrac{\partial}{\partial X_{kl}} \Sigma_i e_{ii}}{\dfrac{V^G}{V^S} - 1 + \Sigma X_{ij} \dfrac{\partial e_{ij}}{\partial X_{mm}} - X_{mm} \dfrac{\partial}{\partial X_{mm}} \Sigma_i e_{ii}}. \qquad \ldots\ldots(2a)$$

If $kl = mm$, we have

$$\frac{\partial p_{mm}}{\partial X_{mm}} = \frac{\Sigma_{ij} X_{ij} \dfrac{\partial e_{ij}}{\partial X_{mm}} - X_{mm} \dfrac{\partial}{\partial X_{mm}} \Sigma_i e_{ii} - 1}{\dfrac{V^G}{V^S}}. \qquad \ldots\ldots(2b)$$

Here V^G and V^S are the specific volumes of gas and solid respectively.

The above expressions give the partial change of vapour pressure of the surface whose normal is the direction mm, when surfaces normal to other axes are not acted on by vapour pressure. To evaluate these expressions accurately it is necessary to know the elastic constants and equations of state of gas and solid over the whole stress range. In the absence of this knowledge the best that can be done is to assume the elastic constants invariable, and to use the values obtained by experiment at low stresses and pressures. For the simple case of unidirectional tension acting on an isotropic solid, and considering evaporation off a stress-free face, these expressions give, for a vapour which is a perfect gas, and whose pressure is small compared with X,

$$\ln \frac{p}{p_0} \approx \frac{V^S X^2}{2E\,RT}, \qquad \ldots\ldots(3)$$

where p and p_0 are the vapour pressures at stress X and at zero stress respectively, and E is Young's modulus. If a value of the stress-free vapour pressure and of the latent heat of evaporation are known at a high temperature, those at any other temperature may be estimated by using the Clausius-Clapeyron equation and the equation giving the temperature gradient of latent heat. The latent heat of a reversible change is the product of the temperature and the change in entropy on passing from one state to another. The stress coefficient of entropy can be expressed in terms of derivatives of stress and strain by making use of the fact that some thermodynamic functions are total differentials, and that for such the order of differentiation is irrelevent. The process for non-hydrostatic systems is formally similar to that for hydrostatic stress. The result is

$$\frac{\partial S}{\partial X_{kl}} = V\left[\frac{\partial e_{kl}}{\partial T} + \left(\Sigma X_{ij} \frac{\partial e_{ij}}{\partial T}\right)\left(\Sigma \frac{\partial e_{ii}}{\partial X_{kl}}\right) - \left(\Sigma \frac{\partial e_{ii}}{\partial T}\right)\left(\Sigma X_{ij} \frac{\partial e_{ij}}{\partial X_{kl}}\right)\right]. \qquad \ldots\ldots(4)$$

As the stress varies, the equilibrium vapour pressure varies, so that for the solid

$$\frac{\partial S}{\partial X_{kl\,(p\ \text{varying})}} = \frac{\partial S}{\partial X_{kl\,(p\ \text{constant})}} + \frac{\partial S}{\partial p}\frac{\partial p}{\partial X_{kl}}. \qquad \ldots\ldots(5)$$

For the gas, the entropy change is simply

$$\frac{\partial S}{\partial X_{kl}} = \frac{\partial S}{\partial p}\frac{\partial p}{\partial X_{kl}}. \qquad \ldots\ldots(6)$$

For any particular stress system, equations (2), (5) and (6) can be solved by graphical methods and the intensity of stress at which L is zero, i.e. the stress at which S is the same for vapour and gas, can be estimated. In the absence of more precise information, Hooke's law and constancy of coefficient of expansion can be assumed. When making the calculation for silica glass in simple tension, it was found that the specific volume of the vapour at high stress was so small that the assumption of a perfect gas was clearly inadmissible; but as the stress varies as the square root of the logarithm of the vapour pressure, the stress calculated on this assumption should not be greatly in error. For practical purposes the maximum stress is reached when the vapour pressure is so high that the solid rapidly disappears by evaporation. For silica glass the maximum stress is estimated to be about $\pm 8\times 10^6$ lb./sq. in. It has already been mentioned that the stress at which the latent heat of evaporation is zero exceeds the breaking stress by a factor of the order of two, and the assumption of invariance of the various coefficients leads to further error. The figure obtained corresponds to a breaking strain of 80%, and it is over twice the strength of fine silica fibres obtained experimentally by Anderegg (1939).

§3. GRIFFITH'S CRITERION AND DELAYED FRACTURE

Griffith considered the equilibrium, when stressed, of a brittle material containing cracks. Consider a rod of such material placed vertically, supported at its upper end, and supporting a load at its lower end. If an initial crack in the material were to spread a little, the load would fall a little, the strain energy of the material would increase slightly, and the total surface energy of the crack surface would increase, due to the increase in area. The loss of potential energy of the load and the gain of strain energy for a given increment in crack length are proportional to the length of the crack, while the gain in surface energy is independent of crack length. For very small crack lengths, the gain in surface energy is greater than the difference between the loss of potential energy and the gain in strain energy, and the crack therefore does not tend to spread. Above a certain crack length the reverse is true, and at a critical crack length, for a given load, the difference between the loss of potential energy and the gain in strain energy produced by an increment in crack length just equals the gain in surface energy. Griffith (1920) considered that at this crack length failure would be immediate, because "the system can pass from the unbroken to the broken condition by a process involving a continuous decrease in potential energy". In a given material with cracks of given length the criterion gives the load at which fracture occurs. Griffith therefore considered that fracture is immediate if spreading of the crack decreases the free energy of the system (all the energies considered by him are free energies). From the quotations from his paper it is clear that he is considering materials to be composed of continuous elastic media, and, applied to such, his critical load would produce immediate fracture. The position is different however when the atomic constitution of materials is introduced. For such materials, even though a state of lower free energy exists, the transition to this state takes time on account of the energy barriers which have to be overcome. For example, the free energy of the system comprising a rod in tension supporting a weight would be reduced if the rod flowed and increased its length while reducing its cross-sectional area, but it is a matter of common experience that, in many materials, the rate of flow is so slow as to be negligible at room temperature. This is because the energy of the average thermal motion is small compared with that of the average energy barrier, and only very rarely does the kinetic energy of atoms exceed that of the average by a quantity sufficient to overcome energy barriers. Other examples of similar delay in attainment of equilibrium are the time taken for materials surrounded by an unsaturated atmosphere to evaporate, and the time for viscous materials, such as glass, to crystallize when maintained at temperatures below the freezing point. The reduction of free energy as a crack spreads is a necessary condition for its spreading, but, from the examples quoted, it is clear that this is not the condition for the crack to spread catastrophically. The rate of spreading will depend on the rate at which thermal motions overcome energy barriers; if much more than the average thermal motion is necessary, the rate of crack-spreading may be negligible. The crack will spread catastrophically only when the average thermal motion is sufficient to overcome the average energy barrier, that is when the stress at the end of the crack equals the maximum the material can withstand.

A possible explanation of the decay in strength of glass is, therefore, the gradual spread of cracks by the material splitting in the way suggested by Griffith.

The thermal energy of mineral glass at room temperature is of the order of a few per cent of the latent heat of vaporization at room temperature, and, therefore, is probably not more than ten per cent of the energy necessary to cause fracture. It seems, therefore, that crack-spreading by splitting due to atoms occasionally possessing more than the average thermal motion is not likely to be effective at stresses of the order of one-third of those necessary to break the material quickly, as in this case the thermal motion would need to be of the order of ninety per cent of the energy to fracture, that is about nine times the average thermal motion. In addition, a certain amount of co-operation between neighbouring atoms would be necessary. If only one atom at a time acquired sufficient thermal energy to cause splitting, the increase in the loads withstood by neighbouring atoms would not be great, and when the atom had lost its excess thermal motion it would probably return to its previous position of equilibrium and no permanent increase in crack length would result. The probability of sufficient co-operative action of neighbouring atoms is likely to be very small. A quantitative estimate of the rate of crack-spreading by the process described above is probably impossible with our present limited physical knowledge, but it is fairly certain that this process would lead to a much slower decay in strength than the 3 : 1 strength ratio corresponding to a $10^6 : 1$ time ratio found in practice.

§4. DELAYED FRACTURE CAUSED BY APPROACH TO HOMOGENEOUS EQUILIBRIUM

Natural processes tend to a reduction in free energy. A stressed solid can attain a state of lower free energy in a number of different ways, and all are possible factors causing delayed fracture. A possible time effect is connected with the entropy change associated with change in stress. Expressions for the stress coefficient of entropy have already been given. The entropy of a solid can be divided into two parts, a part associated with the vibrations of the atoms and a part associated with their mean positions. Change in the kinetic part may be expected to follow change in stress without any appreciable time lag but, on account of energy barriers, the time for atoms to take up their new equilibrium arrangement may be appreciable. The strength of the material will depend on the atomic arrangement, and so it may change with duration of loading. For unidirectional stress X, the entropy change in an isotropic body is given by

$$\frac{\partial S}{\partial X_T} = V^s \frac{\partial e}{\partial T_X}\left(1 - X \frac{\partial}{\partial X_T}\,(2e_x - \overline{e_y + e_z})\right), \qquad \ldots\ldots(7)$$

where $\partial e/\partial T_X$ is the change in strain with temperature at constant stress and e_x, e_y and e_z are the strain in the directions of the three axes.

Assuming Hooke's law, and ignoring variation in $\partial e/\partial T$ with stress, this expression can be integrated and gives the maximum entropy change with stress as

$$V^s \frac{\partial e}{\partial T}\frac{G}{2};$$

where G is the shear modulus. If this is compared with the entropy change known greatly to reduce the strength of a material (for example the entropy change during melting of crystalline materials with latent heats of evaporation of the same order as mineral glass), it is found to be a small fraction of it, less than 1%. The entropy change given above is the combined kinetic and positional entropy, so that the positional effect has been over-estimated, unless the two parts of the entropy change have opposite signs. It is concluded that the entropy change due to stress in material in homogeneous thermal equilibrium, when stress-free, is not likely to cause an appreciable delayed-fracture effect. It has been assumed that the transition temperatures of any possible phase changes have not been reduced to room temperature by the applied stresses. Phase changes under stress are considered in §6.

If the material is not in thermal equilibrium, a more appreciable effect may be expected. Solids which have been cooled from the molten state have a higher free energy than they would if they reached the state which is in equilibrium at their actual temperatures, the excess being greater the more rapidly the material is cooled. This is because, due to viscosity, there is not time for all the atoms to reach their stable equilibrium configurations. Mineral glasses, owing to asymmetry of the molecules, have especially high viscosity, and it would be expected that their excess free energy would be relatively high. In such materials there is a tendency for atoms to rearrange themselves, and in a non-crystalline material some atoms are so situated that little energy is required to move them; these can migrate to new positions. As the strength of materials depends on the relative positions and motions of their atoms, their strength, even if stored stress-free, will vary with age. This is found to be the case for freshly made glass rods and fibres. The approach to thermal equilibrium, however, becomes slower as the least stable atoms find positions of greater stability, and after a time the process effectively ceases and strength becomes independent of age. The atomic arrangement is still not that in equilibrium at the temperature, but the energy required to move the least stable atoms from their immediate environment has become large, compared with the average kinetic energy of the atoms, so that migrations are relatively infrequent. When stress is applied, the average work to enable an atom to migrate is reduced, and the kinetic energy of atoms with energy little more than the average is again sufficient to cause migration at an appreciable rate. By increasing the mobility of atoms, stress thus enables the process of approach to equilibrium to continue, and because the strength depends on the atomic arrangement, strength will vary with duration of loading. The increased mobility of atoms under stress may be looked upon as a reduction in internal viscosity.

An indirect effect of the approach to equilibrium is probably very important. Changes in free energy are accompanied by volume changes, and if such volume changes are non-uniform, internal stresses result. A likely cause of the cracks which are the source of weakness of glass is that they are due to tensile stresses set up by non-uniform volume changes in the glass as its atoms rearrange themselves so as to approach equilibrium configuration. The disintegration of a material during crystallization is the extreme form of the same effect. When a material containing initial cracks is stressed, the approach to equilibrium made possible by the stress will be much more rapid in the highly stressed material at the bottom of a crack than elsewhere. The relatively rapid approach to equilibrium in the material at the bottom of the crack will cause differential shrinkage of this material and will increase the stress. It is possible that this process causes an appreciable part of the delayed-fracture effect in glass.

§ 5. DELAYED FRACTURE CAUSED BY APPROACH TO HETEROGENEOUS EQUILIBRIUM

Under this heading two effects will be considered: delayed fracture caused by evaporation of material at the end of the cracks, and the effects of approach to heterogeneous equilibrium with atmospheric components. Evaporation will be considered first.

If data are available for some high temperature, the vapour pressure of a material at room temperature can be calculated from the Clausius-Clapeyron formula and the equation connecting latent heat with temperature. It is also necessary to know ΔC_p (the specific heat change on vaporization) over the temperature range. If the specific volume and coefficient of expansion of the solid are neglected compared with the corresponding quantities for the gas, the vapour pressures at temperatures T_1 and T_2 are related by the formula

$$\boldsymbol{R}\ln\frac{p_2}{p_1}=\int_{T_1}^{T_2}\frac{\left(L_2-\int_{T_1}^{T_2}\Delta C_p\,dT\right)}{T^2}\,dT, \qquad \ldots\ldots(8)$$

where L is the latent heat.

Using thermal data at 2000° c. for silica glass and assuming the specific heat of silica gas to be the same as that of CO_2 which has the same number of degrees of primary vibrational freedom as SiO_2 molecules, the vapour pressure at room temperature is estimated from equation (8) to be 10^{-78} dynes/cm^2. The vapour pressure is thus entirely negligible. Due to curvature, the stress-free vapour pressure at the end of a crack will be less than that in equilibrium with a flat surface. If the cracks are elliptical, and if the radius of curvature at the end of the crack is assumed to remain constant as the crack spreads, the effect of curvature on vapour pressure can be shown to be equal to that due to a stress equal to the radial component of the surface tension. If the radius of curvature at the ends of the crack is of the order of molecular dimensions, the vapour pressure of silica glass is further reduced by a factor of the order of 10^{-7}, giving a stress-free vapour pressure at the end of the crack of about 10^{-85} dynes/cm^2. However, on account of the logarithmic relationship between vapour pressure and the square of the stress (see equation (3)), at high stresses such as occur at the ends of cracks, the vapour pressure, and consequently the rate of evaporation, may become appreciable. Equation (3) is based on the equilibrium condition in which evaporation and condensation take place reversibly. Before attempting to calculate the rate of evaporation, the reversibility of evaporation and condensation from a stressed solid needs further examination. Under a hydrostatic stress system, the condensed material must withstand the full pressure, and conditions are thus truly reversible, but in the case of a surface having a tension stress in its plane, material condensing on the surface could deposit so as to be stress-free. In the case of a crystalline material strained to a small extent, it is possible that the deposited material would be acted on by the full stress. On account of the difference in interatomic spacing between stressed and unstressed material, if the deposited material is unstressed its atomic arrangement must be somewhat irregular, and this irregularity may entail a higher free energy than if the deposited material withstood the full stress. As the stress is increased, however, the free energy of an amorphous atomic arrangement will become less than that of the stressed crystalline arrangement. It would thus be expected that an unstressed non-crystalline layer would cover the surface of the material. In an amorphous material it would be expected that the deposited layer would be unstressed whatever stress was in the parent material. If, therefore, a stressed material is allowed to remain in contact with its saturated vapour, its surface may be expected to become contaminated with unstressed material and consequently the vapour pressure will fall. If the radius of curvature at the end of the crack in a stressed solid is of the order of molecular dimensions, the highly stressed area will be small compared with the mean free path of the molecules of the gas (except at very high vapour pressures), and as the highly stressed area is surrounded by comparatively unstressed material, with negligible vapour pressure, all atoms escaping from it will condense on the surrounding surface and the material at the end of the crack may be expected to be uncontaminated. The vapour pressure calculated on the basis of a fully stressed surface may therefore be expected to give the correct rate of evaporation from a surface at the end of a crack. If $2a$ is the length of the crack, the rate of spread of the crack due to evaporation estimated from kinetic theory is

$$\frac{da}{dt}=\frac{V^S}{V^G}\frac{\bar{c}}{4}, \qquad \ldots\ldots(9)$$

where V^S and V^G are the specific volumes of solid and gas and $\bar{c}$ is the root mean square of the velocity of the gas molecules, which for a perfect gas at a given temperature can be estimated from kinetic theory. V^G depends on the vapour pressure, which in turn depends on the stress at the root of the crack. This stress for a given average stress can be expressed in terms of the shape of the crack (in principle at least) if the shape of the crack is known. Failure will occur when the crack attains a length such that the stress at its root is the maximum the material can withstand. In practice, failure would be virtually immediate if the vapour pressure became high. Assuming the cracks to be elliptical, the stress at their ends can be expressed in terms of the average stresses, using Inglis's (1913) solution. As the ratio of the major to minor axis of the elliptic crack is large, the stress is sufficiently accurately given by

$$X=2f\sqrt{\frac{a}{r}}, \qquad \ldots\ldots(10)$$

where X is the maximum stress at the end of the crack, f is the average stress to which the material is subjected, $2a$ is the crack length, and r the radius of curvature at the end of the crack. Using equation (3) to give vapour pressure in terms of stress, and assuming the vapour to be a perfect gas, equation (9) becomes

$$\frac{da}{dt}=\frac{V^S}{\sqrt{2\pi RTM}}\,p_0 e^{\frac{2f^2V^S a}{ERTr}}. \qquad \ldots\ldots(11)$$

Integration gives

$$\frac{1}{p}-\frac{1}{p_i}=-\frac{2f^2(V^S)^2}{ERTr\sqrt{2\pi RTM}}\,t, \qquad \ldots\ldots(12)$$

where p_i is the vapour pressure at the end of the crack when the average stress f has just been applied. Failure would occur when, due to crack-spreading, the stress at the end of the crack reaches the maximum the material can withstand. If this is estimated by the methods outlined in § 2, the latent heat of evaporation does not become zero until the vapour pressure is high. At failure, therefore, $1/p$ in equation (12) may be neglected compared with $1/p_i$, and this equation thus leads to

$$\left(\frac{A}{f_0{}^2t_0}\right)^{\frac{1}{f_0{}^2}}=\left(\frac{A}{f^2t}\right)^{\frac{1}{f^2}}, \qquad \ldots\ldots(13)$$

where f and t are the average stress and time to fracture in terms of f_0 and t_0, which are corresponding quantities in an arbitrary basic state. The constant A is given by

$$A=\frac{ERTr\sqrt{2\pi RTM}}{2(V^S)^2p_0}. \qquad \ldots\ldots(14)$$

Assuming r to be the cube root of the volume occupied by a molecule, equation (13) has been applied to silica glass (molecule $=(SiO_2)_1$) having an average strength of about 13,000 lb./sq. in. for a ten-second duration of loading.

If the delayed fracture of this material is wholly attributed to spread of cracks by evaporation and the given values of f_0 and t_0 are put in equation (13), the times to fracture at other stresses can be computed. For example, the time to fracture at a stress 5% less than that giving fracture in 10 sec. is calculated to be of the order of sixty years, whereas the corresponding experimental time is about 30 sec. If only a part of the delayed-fracture effect is attributed to evaporation, the corresponding calculated fracture time will be increased. It seems reasonable, therefore, even having regard to the assumptions and approximations introduced into the calculations, to conclude that the time effect for mineral glass stressed in tension, attributable to evaporation, is very small.

A much more important cause of delayed fracture is atmospheric attack of the glass surface. The decay of ancient stained glass is evidence that this takes place even at room temperature. At higher temperatures glass is readily soluble in water. Barus (1891), when measuring the compressibility of water at elevated temperature, found that the inside of his capillary tube (made of lead glass) rapidly dissolved in water at 185° C. At 350° C. even pure silica glass is readily soluble. The present author, in collaboration with Pearson, has recently investigated the delayed bending fracture of round soda-glass rods of $\frac{1}{4}$ in. diam. when sealed in flexible metal evacuated tubes. The slope of the curve of (stress v. log time to fracture) obtained for the material tested *in vacuo* was about half that for material tested in air at 75% R.H.*

Preston (1942) and Orowan (1944) had previously attributed delayed fracture to atmospheric attack, the former from general and the latter from theoretical considerations. Preston did not give any detailed explanation of the process. Orowan's theory has already been discussed. Instead of attributing delayed fracture to the reduction of surface tension caused by atmospheric attack, it seems better to attribute it directly to loss in strength of the material at the root of the crack. Due to the concentration of strain energy, this material has a much higher free energy than normal unstressed glass and is, therefore, much more chemically active. Continuous reaction with atmospheric constituents is possible even without solution, if the stress at the base of the crack is greater than the strength of the glass-atmospheric constituents complex formed there; for in this case the crack will extend continuously and expose uncontaminated glass to the atmosphere. It is however convenient, and possibly not inaccurate, to regard the process as one of solution of the glass in atmospheric constituents, because in this case the expression already obtained for the delayed fracture caused by evaporation may, with altered constants, be applicable, as solubility in dilute solutions is known to be proportional to vapour pressure. The theory involved in deducing equation (13) gives, for the length of crack (a) in material subjected to constant load, an expression of the form

$$a = -B \log (C - Dt). \qquad \ldots\ldots(15)$$

For crack-spreading by evaporation the constants C and D are very small numbers (of the order of 10^{-40}). For example, if the time to break was 10^5 sec., C might be 10^{-40} and D 10^{-45}. After 10^5 sec., the crack length would be infinite, but after 9×10^4 sec. the crack length would have increased by only $2\frac{1}{2}$%. The crack length is thus effectively constant until a critical time is reached at which it increases extremely rapidly. Crack-spreading by solution would be expected to be much more rapid than crack-spreading by evaporation, and if an equation similar to expression (15) holds, the constants C and D would be much greater, but the crack length may still remain sensibly constant until just before fracture, so that crack-spreading in mineral glass by solution in atmospheric constituents may be consistent with the results of Murgatroyd and Sykes quoted in §1. It is of interest to mention here that rock salt has been found to be appreciably stronger when broken in water (Joffé, 1935). A likely explanation is that, on account of the high solubility, even when unstressed, the highly stressed material at the ends of the cracks becomes covered with a layer of unstressed material deposited from solution. If this happens, on account of the high curvature at the ends of the cracks, the vapour pressure, and hence solubility, will be less than elsewhere, and there will be a tendency for material to deposit at the ends of the cracks and thus to heal them.

---

* In recently published work (1946) Baker and Preston found in some cases that when tests were made *in vacuo* no delayed fracture occurred during the time range 10^{-2} to 10 sec.

§6. THE EFFECT OF POSSIBLE PHASE CHANGES UNDER STRESS

At the beginning of §4, effects due to the entropy change associated with change of stress were discussed, and in materials in thermal equilibrium when stress-free the effects were considered to be unimportant. Much more important entropy changes may occur if, due to stress, the transition temperatures of phase changes is reduced to that of the test temperature, so that phase changes tend to occur. The two phases will be referred to as the high-temperature phase (H.T.P.) and the low-temperature phase (L.T.P.). The sign of the latent heat of transition is taken from the sign of the entropy difference between the H.T.P. and the L.T.P.

The stress coefficient of transition temperature can be calculated by the process used to calculate the stress coefficient of vapour pressure. The result depends on the stress system withstood by the two phases. If the high-temperature phase is a mobile liquid, it can withstand no stress other than hydrostatic pressure. One case to be considered, therefore, is that of a stress-free H.T.P. If both phases can withstand stress, theoretical conditions can be devised so that the relative stresses in the two phases have any given values, but for simplicity the other case treated here is that in which the two phases have the same stresses.

If the H.T.P. is unstressed and the transition takes place on the *mm* face of a cube, and if $kl \neq mm$, the stress coefficient of transition temperature is

$$\frac{\partial T}{\partial X_{kl}} = \frac{\Sigma X_{ij} \frac{\partial e_{ij}}{\partial X_{kl}} - X_{mm} \Sigma_{ii} \frac{\partial e_{ii}}{\partial X_{kl}}}{-\frac{L}{TV^s} - \Sigma X_{ij} \frac{\partial e_{ij}}{\partial T} + \frac{X_{mm}}{V^s} \frac{\partial V^s}{\partial T}}. \qquad \ldots\ldots(16)$$

If $kl = mm$ we have

$$\frac{\partial T}{\partial X_{mm}} = \frac{-1 + \Sigma X_{ij} \frac{\partial e_{ij}}{\partial X_{mm}} - X_{mm} \Sigma_{ii} \frac{\partial e_{ii}}{\partial X_{mm}}}{-\frac{L}{TV^s} - \Sigma X_{ij} \frac{\partial e_{ij}}{\partial T} + \frac{X_{mm}}{V^s} \frac{\partial V^s}{\partial T}}. \qquad \ldots\ldots(17)$$

If the H.T.P. carries the same stress as the L.T.P. we have for $kl \neq mm$

$$\frac{\partial T}{\partial X_{kl}} = \frac{\Delta\left[V \Sigma X_{ij} \frac{\partial e_{ij}}{\partial X_{kl}} - X_{mm} \frac{\partial V}{\partial X_{kl}}\right]}{\frac{L}{T} \pm \Delta\left[X_{mm} \frac{\partial V}{\partial T} - V \Sigma X_{ij} \frac{\partial e_{ij}}{\partial T}\right]}, \qquad \ldots\ldots(18)$$

and for $kl = mm$

$$\frac{\partial T}{\partial X_{mm}} = \frac{\Delta\left[V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{mm}} - X_{mm}\frac{\partial V}{\partial X_{mm}} - V\right]}{\frac{L}{T} + \Delta\left[X_{mm}\frac{\partial V}{\partial T} - V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right]}, \qquad \ldots\ldots(19)$$

where the symbol Δ indicates the difference between the corresponding quantities in the H.T.P. and L.T.P.

For unidirectional stress in an isotropic material, the cases for the stress-free H.T.P. reduce to

$$\frac{\partial T}{\partial X} = \frac{X/E}{-\frac{L}{TV^s} - \frac{X\partial e}{\partial T}}, \qquad \ldots\ldots(16\,a)$$

$$\frac{\partial T}{\partial X} = \frac{1 - 2\nu X/E}{-2X\frac{\partial e}{\partial T} + \frac{L}{TV^s}}. \qquad \ldots\ldots(17\,a)$$

For the stressed H.T.P. case, the equations for unidirectional stress are

$$\frac{\partial T}{\partial X} = \frac{\Delta\left[V\frac{X}{E}\right]}{\frac{L}{T} - \Delta\left[VX\frac{\partial e}{\partial T}\right]}, \qquad \ldots\ldots(18\,a)$$

$$\frac{\partial T}{\partial X} = \frac{\Delta\left[2X\nu\frac{V}{E} - V\right]}{\frac{L}{T} + 2X\Delta\left[V\frac{\partial e}{\partial T}\right]}. \qquad \ldots\ldots(19\,a)$$

In these equations, the latent heat, Young's modulus, specific volume and coefficient of expansion are functions of stress and temperature. The change in latent heat of melting with stress can be derived by methods similar to those used to compute the stress coefficient of the latent heat of evaporation.

It can be seen that at low stresses the stress coefficient of transition temperature is much greater for transition off the face normal to the particular stress than for transition off other faces. The case represented by equations (17) and (17 *a*), however, is of no interest in connection with tensile stresses, as the stress must continue to be applied to the surface of the solid as this recedes, due to melting, and the liquid must remain stress-free; but it has been suggested by Goranson (1940) that it is important in connection with the flow of materials in compression, flow being due to melting under stress. For aluminium a compression stress of the order of 65 tons/sq. in. would reduce the melting temperature to room temperature. Considering the other cases, the stress-free H.T.P. forming off an unstressed face (equation (16 *a*)) leads to a reduction in transition temperature with tension or compressive stress. The stressed H.T.P., forming off a stress-free face (equation (18 *a*)) leads to an increase in transition temperature for tension or compressive stress as $\Delta[V/E]$ is positive for most materials. The stressed H.T.P. forming off the stressed face (equation (19 *a*)) leads to a reduction in transition temperature for tension stress and an increase for compression stress. An idea of the magnitude of the reduction in transition temperature given by equation (19 *a*) may be obtained by computing the temperature coefficient of stress at stresses small compared with E. At low stress the equation reduces to

$$\frac{\partial T}{\partial X} = -\frac{T\Delta V}{L}. \qquad \ldots\ldots(19b)$$

If this formula is applied to a crystalline material, with about the same stress-free "melting" temperature and the same Young's modulus as soda glass, for example aluminium, the estimated stress at which the melting temperature equals room temperature is about 350 tons/sq. in.

It is seen, therefore, that for solids initially in thermal equilibrium, high stresses such as occur at points of stress concentration may be sufficient to cause changes in the stable phase. The phase changes will in many cases take place gradually, and could therefore cause delayed fracture.

Mineral glass at room temperature is not in thermal equilibrium, the high-temperature phase persisting at temperatures below the transition point. The effect of stress on change in the transition point may not therefore be important. If stress lowers the transition temperature, its direct effect would be to reduce the tendency for the glass to crystallize, although the indirect effect (dealt with in §4) of reducing viscosity and so assisting the glass to crystallize may be the more important effect. If stress raises the transition temperature, the free energy reduction on crystallizing will be increased, and for this reason the rate of crystallization will increase, but this would not be expected to be an important direct effect.

The effect of change in transition temperature caused by stress is not, therefore, expected to be important for glass, although it may be important for metals and other materials whose liquid phases have relatively low viscosity. An unknown factor in this connection is the rate of phase transition. In some cases it may be so slow that the transitions are indefinitely delayed.

§7. CONCLUSIONS

1. Previous attempts to explain delayed fracture in glass are not consistent with all known facts.

2. In materials having atomic constitution, Griffith's criterion does not give the stress at which immediate fracture occurs. It gives the lowest stress at which a crack in a material in homogeneous thermal equilibrium may start to spread by a process of splitting. The rate of crack-spreading depends on the rate at which thermal motions overcome energy barriers, but this is considered to be too slow to cause an appreciable delayed-fracture effect.

3. Immediate fracture does not occur until the stress at the end of a crack equals the maximum the material can withstand. The maximum stress is that at which the average thermal motion just overcomes the average energy barrier. The nearest estimate of this stress obtainable from thermal data is that it is the intensity of the applied stress system at which the latent heat of evaporation is zero.

4. In material in homogeneous equilibrium when stress-free, entropy changes due to stress are not likely to cause appreciable delayed-fracture effects.

5. Glass is not in homogeneous thermal equilibrium. Immediately after manufacture, glass tends to approach equilibrium, but the process soon effectively ceases because of the high internal viscosity. Stress reduces the internal viscosity and thus enables the approach to equilibrium to continue. Changes will take place much faster in the highly stressed material at the root of the crack than elsewhere, and as approach to equilibrium results in volume shrinkage, the tensile stress at the ends of cracks increases and causes them to spread. This process is expected to result in an appreciable delayed-fracture effect.

6. Crack-spreading by evaporation of the material at ends of cracks is too slow to give an important delayed fracture effect.

7. Atmospheric attack causes an important delayed-fracture effect. Instead of regarding the effect as due to reduction in surface tension, it is better to attribute it to weakening of the material at the end of the crack. The free energy of this material is much greater than that of normal glass on account of the strain energy concentration, and thus the material is more chemically active. A complex of glass and atmospheric constituents will be formed at the end of the crack. The crack will extend continually if the strength of this complex during or after its formation is less than the load imposed on it.

8. Change, due to stress, in the stable phase at room temperature is not likely to be important for glass, as in this material the high-temperature phase persists at temperatures below the transition point. Phase changes under stress may have important effects on the behaviour of other materials.

LIST OF SYMBOLS

X	Unidirectional stress.
X_{ij}, X_{kl}	Generalized stress.
e	Unidirectional strain.
e_{ij}, e_{kl}	Generalized strain.
p	Pressure.
V	Volume.
V^G, V^L, V^S	Specific volumes of gas, liquid and solid.
E	Young's Modulus.
G	Shear modulus. Also superscript for gas.
S	Entropy. Also superscript for solid.
T	Absolute temperature.
L	Latent heat of isothermal transition.
R	Gas constant.
C_p	Specific heat at constant pressure.
Δ	Difference.
a	Semi-crack length.
$\bar{c}$	Root mean square of velocity of gas molecules.
r	Radius at rooted crack.
f	Average stress.
M	Molecular weight.

REFERENCES

ANDEREGG, F. O., 1939. *Ind. Eng. Chem.*, **31,** 290.
BARUS, C., 1891. *Amer. J. Sci.*, 41, 110.
GORANSON, R. W., 1940. *J. Chem. Phys.*, **8,** 323.
GRENET, L., 1899. *Bull. Soc. Encouragement*, **4,** 839.
GRIFFITH, A. A., 1920. *Phil. Trans.*, **221,** 163.
GURNEY, C. and PEARSON, S., 1945. Awaiting publication.
HOLLAND, A. J. and TURNER, W. E. S., 1940. *J. Soc. Glass Tech.*, **24,** 46.
HOUWINK, R., 1937. *Elasticity Plasticity and Structure of Matter* (Cambridge: The University Press).

Inglis, C. E., 1913. *Proc. Inst. Naval Architects*, **45**, 219.
Joffé, A., 1935. *Rep. Int. Conf. Physics* (London : Phys. Soc.).
Murgatroyd, J. B., 1944. *J. Soc. Glass Tech.* **28**, 406.
Murgatroyd, J. B. and Sykes, F. S., 1945. *Nature, Lond.*, **156**, 716.
Orowan, E., 1944. *Nature, Lond.*, **154**, 341.
Preston, F. W., 1942. *J. Appl. Phys.*, **13**, 623.
Preston, F. W., Baker, T. C. and Glathart, J. L., 1946. *J. Appl. Phys.*, **17**, 162.

CAVITY RESONATORS FOR MEASUREMENTS WITH CENTIMETRE ELECTROMAGNETIC WAVES

By B. BLEANEY, J. H. N. LOUBSER and R. P. PENROSE,
Clarendon Laboratory, Oxford

MS. received 12 *June* 1946 ; *in revised form* 10 *October* 1946

ABSTRACT. A wave-meter for wave-lengths of about a centimetre, with an accuracy of 1 to 2 parts in 10 000, is described, based on a new system of coupling to the H_0 mode in resonant cavities which avoids the excitation of other modes. An indirect effect due to simultaneous resonance in two different modes is discussed in relation to the measurement of absorption by resonant cavities.

The dielectric constants of six non-polar liquids have been determined, by means of wave-meters of this type, at wave-lengths of 3·2 and 1·35 cm. The values obtained at the two wave-lengths agree in all cases within 3 parts in 10 000, which is consistent with the estimated experimental error, and do not differ appreciably from the accepted low-frequency values.

The temperature coefficient of the dielectric constant is compared with that calculated from the dilatation of the fluid. The power factors vary between 10^{-4} and $2\cdot10^{-3}$; the higher values may be due to traces of polar impurity.

PREFACE

The resonant circuit has always played an essential part in apparatus for radio-frequency measurements, and at centimetre wave-lengths its rôle is no less predominant. Its form is, however, very different from the lumped circuit where the electric and magnetic fields are isolated in the capacity and inductance. A hollow resonator is used from which the loss of energy by radiation is negligible, and the dissipation in resistive loss is very small. The ratio of stored energy to energy lost per cycle is therefore large, and the magnification factor Q is high, of the order of 10^4. The resonance is correspondingly sharp, and the wave-length in the cavity can be determined with great accuracy. The resonant cavity is thus eminently suitable for use as a wave-meter, and for the determination of the properties of low-loss dielectrics at centimetre wave-lengths.

The difference between the lumped circuit and the resonant cavity is not only in form but also in nature. The concepts of current and voltage are replaced by those of magnetic and electric field, and the concept of impedance is therefore only of subsidiary importance. In Part I of this paper the theory of a new method of exciting one particular mode of resonance, the H_0, in the cavity, is developed by considering the fields in the cavity and in the wave-guide feeder. A precision wave-meter based on these principles is described, which covers the wave-length band from 1·15 cm. to 1·55 cm.

In Part II the application of such wave-meters to the measurement of the dielectric properties of six non-polar liquids at wave-lengths of 3·2 cm. and 1·35 cm. is described, and the results are briefly discussed.

PART I.—DESIGN OF WAVE-METER

§1. INTRODUCTION

It is convenient to divide hollow resonators into two types: coaxial line resonators and wave-guide (cavity) resonators. In the coaxial line the waves of the normal type travel with the same velocity as in free space, and the wave-length may be directly determined from the distance between successive points of resonance, which will occur at intervals of half a wave-length. To ensure that only waves of the normal type are present, the diameter of the outer conductor of the coaxial line must be small compared with the wave-length. At the shorter wave-lengths this restriction results in considerable mechanical difficulties in the design of coaxial line resonators, and also increases the ratio of the dissipation of energy per cycle to the stored energy. The Q value is therefore smaller and the accuracy of setting is reduced. These difficulties are overcome by the use of wave-guide resonators, with which the problem becomes the elimination of all but the desired wave-type. One simple solution is the use of a wave-guide in which only one mode of propagation, the lowest, or H_1 mode, is possible. The diameter * of the wave-guide must, for this purpose, lie between 0·59 and 0·77 of the wave-length. The useful wave-length range is therefore less than 20%. Moreover, a much more serious difficulty arises at wave-lengths of a centimetre or less. In a wave-guide the velocity of propagation depends on the ratio of the wave-length to the diameter of the guide. The latter must therefore be constant and known with high precision; when the diameter is only a few millimetres the machining tolerances demanded become prohibitive. These difficulties may be overcome by the use of a more suitable type of wave than the H_1.

Although an unlimited number of modes of propagation are available in a wave-guide of sufficient size, the choice of mode for an accurate wave-meter is readily narrowed down to one, the H_0. The especial property of this mode—of giving a low dissipation of energy on the walls of the guide—has long been recognized. In addition it has the following advantages:—

1. There is no flow of current in the radial direction on the end walls. The cavity may therefore be tuned by a non-contact piston without leakage of energy past the piston.

2. The H_0 mode is a singlet mode, whereas all other modes (except E_0) are doublets, corresponding to the possibility of either cosine or sine distribution of field with respect to azimuth. The degeneracy of these doublet modes is removed by a slight ellipticity, and double resonances may appear. The H_0 mode is stable against small irregularities in the guide.

---

* The advantages of circular guide (e.g. in machining) are so obvious that discussion of other shapes is omitted.

3. The diameter of the guide must be greater than 1·22 wave-lengths, and with the coupling system described below, a diameter as large as 2½ wave-lengths has been used. For the shorter wave-lengths this makes the machining tolerances much easier.

A large size of guide permits, of course, the propagation of a number of other wave-types, which may give rise to undesired and misleading resonances. To remove the ambiguities thus caused it has hitherto been customary to attempt to damp-out the undesired resonances by placing " lossy " material (e.g. resistive wires) at suitable positions within the resonator. Such methods are only partially effective, and have the disadvantage of introducing irregularities in the cavity.

A more fundamental approach to the problem suggests the elimination of undesired resonances by the use of a properly designed system of coupling the resonator to the feeder. A coupling unit such as a probe or loop, or, preferably, a hole, can only feed energy into modes of resonance which have the appropriate field components at the coupling unit. Modes in which these components are zero will not be excited. This principle is combined, in the coupling system described below, with the use of a double symmetrical feed which eliminates all modes whose fields vary with $\genfrac{}{}{0pt}{}{\cos}{\sin} n\phi$ round the periphery if n is an odd integer. This elimination depends, of course, on an exact balance between the two halves of the feed, which is obtained by the use of two symmetrical holes of equal diameter. It is for this purpose that holes are greatly to be preferred to probes or loops, since the balance is automatically achieved, whereas probes or loops would require delicate adjustment by hand which would probably be frequency-sensitive.

The correct location of the coupling holes in the feeder is also important. When the feeder is a rectangular H_1 wave-guide, as is usual at the shorter wave-lengths, the obvious position for the coupling holes is on the narrow side of the guide, since the only field component here is a longitudinal magnetic field. If the coupling holes are small, this field will not be seriously disturbed by their presence, and only modes in the resonator which have a parallel magnetic-field component will be excited. On the broad side of the guide there are three field components, and the number of modes likely to be excited in the resonator is correspondingly greater. Thus as a general rule coupling out of this side is to be avoided.

The theory of this system of coupling is developed in the following sections, and two H_0 wave-meters based on these principles are described. Freedom from undesired resonances over a considerable wave-length range is obtained, the performance in this respect fully confirming the predictions.

§2. THEORY OF THE COUPLING SYSTEM

The field components at the end wall of a cylindrical cavity are (cf. Lamont, *Wave-guides* (Methuen), p. 76):

for *H*-waves	for *E*-waves
$H_z=0$	$H_z=0$
$H_\rho=(2\pi k/\lambda_g)J_n'(k\rho)\cos n\phi$	$H_\rho=-j(2\pi n/\lambda_a\rho)\,J_n(k\rho)\sin n\phi$
$H_\phi=-(2\pi n/\lambda_g\rho)J_n(k\rho)\sin n\phi$	$H_\phi=-j(2\pi k/\lambda_a)\,J_n'(k\rho)\cos n\phi$

for *H*-waves (*cont.*)

$$E_z = 0$$
$$E_\rho = 0$$
$$E_\phi = 0$$

for *E*-waves (*cont.*)

$$E_z = k^2 J_n(k\rho) \cos n\phi$$
$$E_\rho = 0$$
$$E_\phi = 0$$

where the nomenclature is the same as that used by Lamont, viz. (z, ρ, ϕ) are cylindrical polar co-ordinates, k is defined by $J_n'(ka) = 0$ where a is the radius of the resonator, and J_n is the Bessel function of order n. λ_g, λ_a are the wave-lengths in the cavity and in free space respectively.

The only field existing along the narrow side of an H_1 rectangular wave-guide is a longitudinal magnetic field. If, therefore, coupling holes in the side wall open into the end of a cylindrical cavity along an axial plane, only the radial component of the magnetic field in the cavity will be coupled to the field in the wave-guide. Thus we need only consider the behaviour of H_ρ in the equations above. Since H_ρ varies as $\cos n\phi$, the radial fields at diametrically opposite points will be of the same sign if n is even, and of opposite sign if n is odd. This is illustrated by the sketches (figure 1) of the magnetic fields H_0, H_1 and H_2. It is possible, therefore, to avoid the excitation of any *E*-wave or *H*-wave of odd order by using two equal coupling holes spaced half a wave-length apart in the rectangular wave-guide, and placed at diametrically opposite points in the cylindrical cavity.

Table 1. Ratio of diameter to cut-off wave-length for the principal modes in a cylindrical resonator

H-waves	H_{n_1}	H_{n_2}	H_{n_3}
H_0	1·2197	2·2331	
H_1	0·5861	1·6970	2·7172
H_2	0·9722	2·1346	3·1734
H_3	1·3373	2·5513	
H_4	1·6926	2·9547	
H_5	2·0421		
H_6	2·3877		
H_7	2·7304		
H_8	3·0709		

E-waves	E_{n_1}	E_{n_2}	E_{n_3}
E_0	0·7655	1·7571	2·7546
E_1	1·2197	2·2331	
E_2	1·6347	2·6793	
E_3	2·0309	3·1070	
E_4	2·4154		
E_5	2·7920		
E_6	3·1628		

***Accuracy*. ±1 in last place (or better). All modes for which d/c is less than 3 are included.**

Table 1 gives values of d/λ_c for various wave types. In a typical H_0 cavity, d/λ_c is 1·52 at the centre of the band; waves for which d/λ_c exceeds this value are therefore beyond cut-off. Of the permissible modes of resonance in this size of resonator, H_1, H_3 and E_1 are not excited for reasons already discussed; neither is E_0, for it has no radial component of H. The only undesired modes which are liable to occur are therefore H_2, H_4 and E_2; of these, the last two do not arise except at the shorter wave-lengths. The coupling to the higher-order waves is in any case weak, since their radial components of magnetic field are small at the coupling holes. In a transmission-type wave-meter it is possible to eliminate the H_2 resonances by suitable placing of the output coupling, since the field components vary with $\sin 2\phi$ or $\cos 2\phi$. It is convenient to use an output coupling of the usual type—a single hole in the side of the wave-meter barrel a quarter wave-length from the end wall—feeding a rectangular wave-guide in which the plane containing the magnetic field passes through the axis of the cylindrical cavity. The output guide then couples to the H_z field in the cavity; for H_2, this field is zero at an angle of 45° to the input guide. Thus if the output is oriented at this angle (as in figure 2) it will not couple to any E-wave, since by definition these have no longitudinal component of magnetic field.

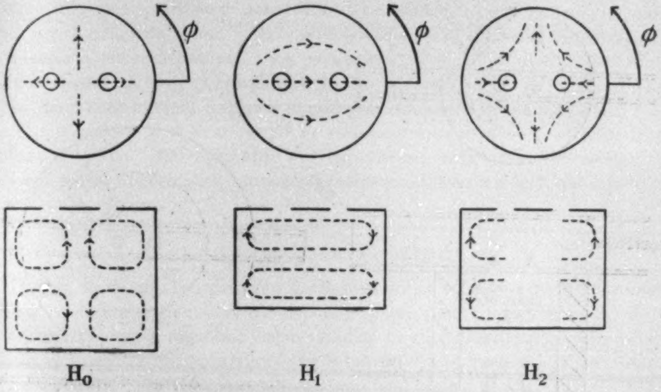


Figure 1. Magnetic fields in cylindrical cavity.

Band-width of the coupling system

Since the input coupling makes use of two holes spaced half a wave-length apart, its performance will not be so good at wave-lengths other than that for which it was designed. The diminution in its efficiency is determined solely by the change of wave-length in the input wave-guide; change of wave-length in the resonant cavity has no effect since the two coupling holes are placed in the end wall. It is easily seen that if l is the distance between the coupling holes, the coupling to waves of order $2n$ in the cavity is proportional to $\sin \pi l/\lambda_g$ and that to waves of order $2n+1$ is proportional to $\cos \pi l/\lambda_g$. Thus when $l=\lambda_g/2$, the coupling to the wanted waves is a maximum while that to the unwanted waves is zero. To find the useful band-width of the device, we take as a somewhat arbitrary criterion that the coupling to the $2n$ modes (as measured by a square-law detector) shall be at least ten times that to the $2n+1$ modes, assuming that the coupling is determined mainly by the sine and cosine terms. We have, then, at the end of the band

$$\tan^2 \pi l/\lambda_g = 10,$$

the solution of which gives approximately $\lambda_g = 2{\cdot}5l$ or $1{\cdot}6l$. Taking the typical value 1·25 for the ratio λ_g/λ at the centre of the wave-band λ_0, one finds that the instrument should work satisfactorily at wave-lengths between $1{\cdot}15\lambda_0$ and $0{\cdot}90\lambda_0$—a 25% band.

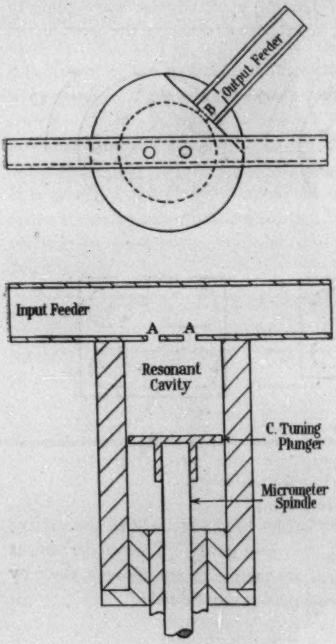


Figure 2. H_0 resonator.

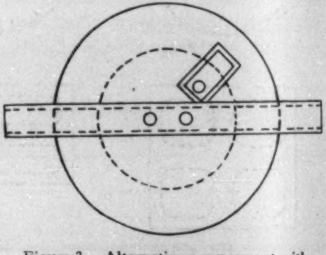


Figure 3. Alternative arrangement with output coupling on end.

§ 3. EXPERIMENTAL CONFIRMATION

The conclusions drawn in the previous paragraph were tested in the region of 1 cm. wave-length. A cylindrical cavity was made with an internal diameter of 19 mm. (figure 2); it was coupled to the H_1 rectangular wave-guide by two holes (AA) spaced 7·7 mm. apart, which in this particular guide corresponds to a half wave-length for a frequency of 24 000 Mc./sec. (1¼ cm. wave-length). The output coupling hole B was drilled in the circular barrel at a point one quarter of a wave-length (in the cavity) from the end, and oriented at 45° to the axis of the input coupling. The cavity was tuned by a plunger C driven by a

$2\frac{1}{2}''$ micrometer head with 25 mm. travel, manufactured by Messrs. Moore and Wright. The coupling holes were adjusted in size by experiment until adequate power was available in the detector, a crystal.

In order to make a direct comparison of the performance of the new coupling system with that of the conventional coupling system, a fourth coupling hole was drilled similar and diametrically opposite to the hole B. This fourth hole and B form a coupling system of the usual type, and the resonator could be quickly transferred between this system and the new coupling system so that they could be compared under identical conditions. This comparison was made at $1\frac{1}{4}$ cm. wave-length. With the conventional coupling system, several modes of resonance were easily detected, one of them, H_3, being over twice as strong as the desired H_0 resonance! On the other hand, with the new coupling system only one undesired mode could be detected by a systematic search. This was the H_2 mode, and it gave a response smaller than that of the H_0 mode by a factor of several hundred. Its presence was probably due to the fact that in the first model the output coupling hole was not quite at 45° to the input holes; in any case, such a small response would pass unnoticed in ordinary use.

The resonator was next tested at various points in a range of wave-length 1·08 to 1·48 cm. No difficulty was experienced with undesired modes; the performance of the coupling system was satisfactory over the 25% band predicted by the considerations of § 2.

§ 4. THE WAVE-METER

During work on absorption in gases near 1 cm. wave-length, the need arose for a wave-meter with an accuracy of one or two parts in ten thousand. In the first experiments, a resonant cavity similar to that described in the previous section was used to measure wave-length, but its accuracy was only about one part in a thousand. The limitation of accuracy was almost entirely due to the size of the barrel. The effect of errors in the diameter of the barrel may be estimated from the equation

$$\frac{d\lambda_c}{\lambda_c}=\frac{d\lambda}{\lambda}\cdot\frac{\lambda_c^2}{\lambda^2} \qquad \ldots\ldots(4.1)$$

obtained by differentiation of the usual wave-guide equation. Here λ_c is the critical wave-length of the cavity ($=0{\cdot}82$ diameters) and λ is the free-space wave-length of the radiation. If $d\lambda/\lambda$ is taken as 10^{-4}, the diameter of the barrel as 19 mm., and $\lambda=1\frac{1}{4}$ cm., the allowable error in the diameter is only 3 microns. With the resources available in this laboratory the diameter could neither be machined nor measured to this accuracy.

The solution of this difficulty is suggested by the form of equation (4.1), in which the allowable error in the diameter depends on the cube of λ_c. If the diameter of the barrel is increased to 30 mm., the tolerance becomes 12 microns, which is within the limits of good instrument making. The disadvantage of using so large a diameter is, of course, that it permits the propagation of many modes (16 at $1\frac{1}{4}$ cm. wave-length). The new coupling system eliminates, by its symmetry and the use of the 45° angle between input and output, all but the modes for which n is a multiple of 4. The fields of the modes with high values of n are concentrated towards the periphery and are relatively weak near the axis, corresponding to the fact that in the expansion of $J_n(x)$ the lowest term is of the order x^n. Thus in a large diameter cavity, the coupling to modes such as E_4 and H_4 is small because the distance of the input coupling holes from the axis of the cavity is only a small fraction of the radius of the cavity.

The drilling of the output coupling hole in the barrel, and the machining of a flat surface on the outside for the attachment of a wave-guide, are liable to cause an undesirable distortion of the barrel. This may conveniently be avoided by placing the output coupling on the end, again at 45° to the input. Such an arrangement is shown in figure 3. The output hole is located at a distance from the axis equal to 0·55 times the radius of the cavity, where the radial overtone (H_{02}) of the desired mode has zero radial magnetic field. This eliminates the overtone, which otherwise would give a resonance of about the same intensity as the fundamental.

A wave-meter fed in this manner, and with a cavity diameter of 30 mm., has been constructed from copper, and has been very satisfactory. It has been used for wave-lengths ranging from 1·15 cm. to 1·55 cm. without any ambiguities from other modes. The plunger is driven by a 2½-inch micrometer head, graduated to 0·002 mm., which is normally read to half a division. The consistency of the measurements obtained with the instrument is illustrated by the two following sets of readings for the position of resonance, where a fourth decimal place has been determined by estimation of fractions of a small division:—

Micrometer reading	Half wave-length	Micrometer reading	Half wave-length
24·1984		23·8437	
	7·3459		7·2573
16·8525		16·5864	
	7·3449		7·2560
9·5076		9·3304	
	7·3440		7·2558
2·1636		2·0746	

The slight increase in the half wave-length at the higher micrometer readings is consistent with a slight taper in the barrel (less than 0·01 mm. in the diameter) which was just observable by direct measurement. If, in general, an error of 0·002 mm. in reading the wave-length in the cavity is allowed, the corresponding error in the free-space wave-length is about 1 part in 10000. From the width of the resonances the Q of the cavity was estimated as 20000; the positions of resonance could therefore easily be located to within 0·001 mm. The error in the measurement of the diameter of the barrel should not exceed 0·01 mm. An overall accuracy of 2 parts in 10000 in the wave-length may therefore be expected.

§5. INDIRECT EFFECT OF OTHER MODES OF RESONANCE

Examination of the performance of the wave-meter showed that at certain wave-lengths one of the H_0 resonances was abnormally small and broad. The other resonances were sharp and of normal intensity. It was found that these abnormal resonances occurred at points where the cavity could be simultaneously resonant in two different modes; for instance, when three half wave-lengths of the H_2 mode are equal to two half wave-lengths of the H_0 mode. Thus at a wave-length of 1·37 cm., the second H_0 response in a cavity of 19 mm. diameter was found to be smaller by a factor ten than the usual value. The adjacent H_0 resonances were, of course, unaffected. The wave-length at which n_0 half wave-lengths of H_0 are equal to n_1 half wave-lengths of another mode is given by the formula

$$\frac{\lambda}{d}=\frac{n_0^2-n_1^2}{n_0^2x_1^2-n_1^2x_0^2},$$

where d is the diameter of the barrel, and x_0 and x_1 are the values of d/λ_c appropriate to H_0 and the other mode respectively (see table 1). Some of the more important points where resonances coincide are given in the following table:—

Table 2

Interfering type	n_0	n_1	λ/d
E_2	1	0	0·612
H_1	1	2	0·733
E_0	1	2	0·749
H_3	2	1	0·730
H_1	2	3	0·644
E_0	2	3	0·672
H_2	2	3	0·722

Since the resonances due to other modes are too small to be detected by themselves, their effect on the H_0 mode at coincidence is presumably due to some asymmetry in the cavity which causes a transfer of power from the H_0 to the undesired mode. An obvious asymmetry is the presence of the coupling holes which disturb the normal current flow within the resonator. Evidence supporting this suggestion was obtained from a "reaction" wave-meter, in which the output coupling was absent and resonance was detected by the absorption of part of the power transmitted along the feeder wave-guide. The only "crossover" point at which the response was low was for two half wave-lengths, H_0 mode, equal to three half wave-lengths, H_2 mode. When a similar experiment was tried with a cavity containing an output hole, the effect of crossover with the H_1 mode could also be observed.

So far as could be ascertained, no shift of the position of resonance occurred at a crossover, and the performance of the cavity as a wave-meter was not affected. The presence of crossovers may, however, cause considerable errors if the cavity is being used to measure absorption or power factor. This is illustrated by an effect observed during use of an evacuated cavity at a wave-length where a crossover occurred, and the response was about ten times smaller than normal. When air was admitted to the cavity the response rose by a factor of three. The slight change in dielectric constant removed the coincidence of the resonances, owing to the different dispersion of the two modes. The difficulties associated with the presence of crossovers make it advisable, in absorption measurements, to use a cavity of the smallest possible diameter and to work on the first or second half wave-length, where crossovers occur infrequently. A check on the presence of a crossover should always be made by observation of several resonant positions.

The effect of these mode crossings may be reduced by damping the unwanted mode. This damping must be effected without introducing more discontinuities in the cavity, since further asymmetry is likely to increase the coupling between the modes and thus enhance rather than diminish the effect of crossovers. A good method utilizes the leakage of all modes except the H_0 past the non-constant plunger; a ring of an absorbent material such as bakelite-bonded paper or linen placed on the back of the plunger is quite effective. In a cavity designed for vacuum work (Bleaney and Penrose, 1946) water contained in an annular glass ring has also been used.

PART II.—DIELECTRIC PROPERTIES OF SIX NON-POLAR LIQUIDS AT WAVE-LENGTHS OF 3·2 CM. AND 1·35 CM.

§ 6. INTRODUCTION

The use of cavity resonators for measuring the dielectric properties of solids and liquids is well known. Penrose (1946) and others have used a cavity excited in the H_0 mode for measurements on solid dielectrics. A specimen of known dimensions is placed in the cavity, and from the shift of the position of resonance and the breadth of the resonance curve the dielectric constant and the power factor can be calculated. A similar method has been used (Whiffen and Thompson, 1946; Horner, Taylor, Dunsmuir, Lamb and Jackson, 1946) for liquids of high power factor. There are, however, several disadvantages in using a cavity only partly filled with liquid:

1. The wave-length in the liquid is not measured directly but deduced from the shift of the resonances in the air-filled portion of the cavity.

2. The height of the liquid in the cavity must be known accurately.

3. The surface of the liquid is not flat because of surface-tension effects.

4. A minor objection is that the remainder of the cavity is filled with the vapour. The error in the dielectric constant arising from this source will be only a few parts in 10 000.

For measurements on liquids with good power factors, these disadvantages may be eliminated by using a cavity completely filled with the liquid. Both the experimental procedure and the mathematics are simplified: thus, for the dielectric constant only two measurements of wave-length are required after the preliminary determination of the diameter of the cavity. In addition, the broadening of the resonance curve due to the presence of an imperfect dielectric is, of course, greatest when the cavity is filled with the dielectric.

The dielectric constant may be calculated from measurements on a filled cavity by means of the simple formula:

$$\left.\begin{aligned} &\text{Air filled cavity:} && \frac{\epsilon_a}{\lambda^2}=\frac{1}{\lambda_a^2}+\frac{1}{\lambda_c^2}. \\ &\text{Liquid filled cavity:} && \frac{\epsilon}{\lambda^2}=\frac{1}{\lambda_d^2}+\frac{1}{\lambda_c^2}, \end{aligned}\right\} \quad \ldots\ldots(6.1)$$

where λ = wave-length in free space,

λ_a = ,, ,, in air filled resonator,

λ_d = ,, ,, in liquid filled resonator,

λ_c = critical wave-length of cavity = diameter/1·2197,

ϵ_a = dielectric constant of air = 1·0006,

ϵ = dielectric constant of liquid.

§7. THE APPARATUS

Resonator A, for 3·2 cm. wave-length (barrel diameter 5·075 ± 0·001 cm.) had a side coupling to the detector, as described in §3. The cavity was made liquid-tight by clamping thin discs of mica over the coupling holes.

A side output has the drawback that its electrical position in the resonator varies with the dielectric constant of the filling. This defect was avoided at 1·35 cm. wave-length by using an output from the end, as described in §4; this arrangement has also the advantage of requiring only one piece of mica to close all the coupling holes. Resonator B had a diameter of $2{\cdot}106_5 \pm 0{\cdot}001$ cm., and to fit the output on to the end, a rectangular guide smaller than the normal was used. This was filled with ebonite to increase its cut-off wave-length beyond 1·35 cm. In addition, the precision wave-meter of §4 (here called resonator "C") was used, as its large diameter makes negligible the error due to uncertainty in the barrel diameter.

When the cavity is filled with liquid, the number of possible modes of resonance is increased. The higher modes leak past the plunger, and can therefore be damped down by placing an absorbent material behind the plunger. A sealed glass tube containing water was used thus in resonator A, as most other "lossy" materials would contaminate the organic liquids under investigation; this proved unnecessary in the two smaller resonators with end output, which seems to be superior in this respect to a side output. The probability of coincident resonances is also increased by the dielectric constant of the liquid, but they could be avoided by slight readjustment of the oscillator frequency. At the shorter wave-length, four resonant points could be obtained in the liquid, and any resonance abnormally broad and weak was discarded.

Drift of the oscillator frequency was negligible at 3·2 cm. wave-length. At 1·35 cm. wave-length a small correction was necessary, determined by successive measurements with the cavity alternately filled with air and with liquid.

§8. RESULTS FOR THE DIELECTRIC CONSTANT

The values for the dielectric constants of the six non-polar liquids at a temperature of 20° c. are shown in table 3.

Table 3. Dielectric constant at 20° c.

Liquid	Purity	Dielectric constant			
		λ=3·2 cm.	λ=1·35 cm.		Low-frequency values
		Resonator "A"	Resonator "B"	Resonator "C"	
Benzene	Commercial	2·2850	2·2853		
,,	Analar	2·2835	2·2828	2·2830	
,,	Analar dried over sodium	2·2780	2·2778	2·2776	2·2825 (*a*)
Cyclo-hexane	" A "	2·0244	2·0246	2·0251	2·020 (*b*)
n-heptane	" A "	1·9220	1·9223	—	1·926 (*c*)
CS_2	" A "	2·6476	2·6477	—	2·630 (*d*)
n-hexane	Commercial	1·9016	1·9016	—	1·919 (*e*)
CCl_4	Commercial	2·2386	2·2390	—	2·2409 (*f*) 2·2445 (*g*)

Notes

" A " Purified sample kindly lent by Dr. H. W. Thompson of the Physical Chemistry Laboratory, Oxford.

(*a*) Hartshorn and Oliver, 1929. *Proc. Roy. Soc.*, **123, 664.** Values ranging from 2·276 to 2·289 have been obtained by different experimenters (see reference (*g*) below).

(*b*) Hooper and Kraus, 1934. *J. Amer. Chem. Soc.*, **46,** 2265.

(*c*) Smith, Morgan and Boyce, 1928. *J. Amer. Chem. Soc.*, **50,** 1883.

(*d*) Wolfke and Mazur, 1932. *Z. Phys.*, **74,** 110.

(*e*) Chrétien, 1931. *C.R. Acad. Sci., Paris*, **192,** 1385.

(*f*) Goss, 1933. *J. Chem. Soc.*, p. 1341.

(*g*) Clay, Dekker and Hemelrijk, 1943. *Physica*, **10,** 768.

In no case do the values obtained in the different resonators differ by more than 0·03%. This is consistent with the possible errors, which are as follows:—

(1) *Error in the barrel diameter*

Variation of the diameter and uncertainty in its measurement amounted to 0·001 cm. From equations (6.1) one finds

$$\frac{\delta\epsilon}{\epsilon}=2\left(\frac{\lambda}{\lambda_c}\right)^2\left(\frac{\epsilon-1}{\epsilon}\right)\frac{\delta\lambda_c}{\lambda_c}. \qquad \ldots\ldots(8.1)$$

The corresponding errors in the dielectric constant are 0·01% in resonators " A " and " C " and 0·02% in " B ".

(2) *Periodic error in the micrometer*

Calibration by the N.P.L. showed that this was not more than 0·0002 mm. The scale was normally read to half a division, or 0·001 mm., and an error of this magnitude corresponds to less than 0·01% in ϵ.

(3) *Frequency drift of the oscillator*

In the worse case ($\lambda = 1{\cdot}35$ cm.) the percentage change in frequency was not more than 0·03% during the experiments on any one liquid, corresponding to 0·02% in ϵ. A correction has been applied as described above.

(4) *Change in temperature*

An error of 0°·2 c. in the temperature would cause an error in ϵ of 0·01%. The resonators were lagged and the temperature measured with a probable error of 0°·2 with a thermometer previously compared with a standard. The liquid in the resonator could be stirred effectively by spinning the plunger rapidly.

§ 9. THE TEMPERATURE COEFFICIENT OF THE DIELECTRIC CONSTANT

By cooling or heating the resonator slightly, the dielectric constant was determined at a number of temperatures within the range (6–30° c.). The temperature coefficient was found from a plot of ϵ against t. A correction was applied for the expansion of the metal parts of the resonator.

The values of $(d\epsilon/dt)_{20^\circ}$ obtained at a wave-length of 3·2 cm. are shown in column (2) of table 4, together with the low-frequency values given in the International Critical Tables. The variation of the dielectric constant with the temperature is principally due to expansion of the liquid. From the Clausius–Mossotti formula one finds

$$\frac{d\epsilon}{dt} = -\frac{(\epsilon-1)(\epsilon+2)}{3}\cdot\gamma,$$

where γ is the coefficient of cubical expansion, assuming that the molar polarizability is independent of temperature. Using the expansion coefficients given in the Landolt–Börnstein tables, the values of $d\epsilon/dt$ have been calculated and are shown in the last column of table 4. They are generally in good agreement with the measured values.

Table 4

Liquid	$\frac{d\epsilon}{dt}$ at 20° c.	$\frac{d\epsilon}{dt}$ at 20° c. (according to I.C. Tables)	$\frac{d\epsilon}{dt}$ calculated from expansion coefficient
Benzene—Commercial	$-{\cdot}0020 \pm {\cdot}0001$	$-{\cdot}00205$	—
,, —Analar	$-{\cdot}0020 \pm {\cdot}0001$	$-{\cdot}00198 \pm {\cdot}00003$	·0020
n-hexane	$-{\cdot}0015 \pm {\cdot}0001$	$-{\cdot}0011$	·0016
Cyclo-hexane	$-{\cdot}0016 \pm {\cdot}0001$	$-{\cdot}0016$	·0016
n-heptane	$-{\cdot}0013 \pm {\cdot}0001$	$-{\cdot}0014$	·0015
Carbon tetrachloride	$-{\cdot}0019 \pm {\cdot}0001$	$-{\cdot}0014$	·0021
Carbon bisulphide	$-{\cdot}0026 \pm {\cdot}0001$	—	·0030

§ 10. THE POWER FACTOR OF THE LIQUIDS

By measurement of the width of the resonance curve of the cavity when filled with liquid, a maximum value of the power factor of the liquid can be obtained. To find the true value it is necessary to correct for the dissipation of energy in the cavity elsewhere than in the dielectric. These other losses are due to the finite conductivity of the walls of the cavity and to the finite size of the coupling holes by which energy is fed into and out of the cavity. By keeping these holes small, it is generally possible to make these coupled losses small (e.g. 10%) compared with the loss of energy on the walls. The latter cannot be measured directly when the liquid is present: it might be calculated assuming the conductivity of the wall to be given by the low-frequency conductivity of the bulk metal. This will certainly give a minimum value for the loss on the walls; in practice the losses are greater than this theoretical value, as is shown by the fact that the measured Q values of empty cavities are generally found to be somewhat lower than the theoretical value at centimetre wave-lengths.

These difficulties in making a correct allowance for losses other than that in the dielectric emphazise the importance of making these losses as small as possible in comparison with that in the dielectric. In a cavity resonant in the H_0 mode, the losses on the wall are small; the correction to the power factor is about 0·0001 in these experiments. From the measured Q value of the empty cavity, an apparent conductivity of the walls may be calculated, assuming the conductivity to be the same everywhere. This apparent conductivity is then used to calculate the energy lost on the walls when the liquid is present, since the current distribution is known once the dielectric constant has been determined. This method is valid when the coupled loss is small compared with the loss on the walls; it is likely to be more inaccurate when the liquid under test has a high dielectric constant, since the change in the current distribution will be large. In these experiments the dielectric constants of the liquids were all about 2, and the coupled loss was about 5% of the wall losses; the correction should therefore be reliable.

The values of the power factors of the liquids at two wave-lengths, 3·2 cm. and 1·35 cm., are shown in table 5. The correction for the losses on the walls of the cavity has been applied; it amounted to $0{\cdot}0001_0$ at 3·2 cm., and $0{\cdot}0001_3$ to $0{\cdot}0001_8$ (depending on the dielectric constant) at 1·35 cm.

Table 5. Power factors of the liquids

Liquid	Purity	Power factor $\lambda=3{\cdot}2$ cm.	 $\lambda=1{\cdot}35$ cm.
Benzene	Commercial	0·00057	0·0017
,,	Analar	0·00050	0·0012
,,	Analar dried over sodium	0·00035	0·00087
Cyclo-hexane	"A"	0·00005	0·00019
n-heptane	"A"	0·00037	0·00076
CS_2	"A"	0·00024	0·00072
n-hexane	Commercial	0·00034	0·00076
CCl_4	Commercial	0·00031	0·00078

§ 11. CONCLUSION

The internal consistency of the measurements of the dielectric constant is very gratifying, and accords with the accuracy expected of the method. The apparatus is simple, and only metrological measurements are required. The differences between these values and those previously obtained by other experimenters at low frequencies are generally somewhat greater than the possible error in the measurements. There is no reason to expect that the dielectric constants at centimetre wave-lengths of these non-polar liquids should be appreciably different from the static values, and the difference may be ascribed to small quantities of impurity. Thus the value for the purified benzene is about 0·2% lower than the generally accepted value of Hartshorn and Oliver (1929), but lies within the range of low-frequency values obtained by various methods between 1922 and 1943.

All the liquids measured show appreciable power factors at these short wave-lengths, being greater by a factor of 2 to 3 at 1·35 cm. than at 3·2 cm. This may be due to polar impurities; from the measurements of Whiffen and Thompson (1946) the necessary concentration of polar impurity can be estimated as a few hundredths of one per cent. The presence of such small amounts cannot be excluded, though it is surprising that the power factors of the pure samples "A" and the dried analar benzene are no better than those of the samples of CCl_4 and *n*-hexane, which were ordinary commercial grade. Only one liquid, cyclohexane, was appreciably better in power factor than the rest; for this liquid the measured power factors scarcely exceed the possible experimental error.

ACKNOWLEDGMENT

The work described in Part I of this paper was carried out under research contract for the Director of Scientific Research, Admiralty, and the authors wish to thank the Board of Admiralty for permission to publish the paper.

REFERENCES (see also notes to table 3)

Bleaney and Penrose, 1946. *Proc. Roy. Soc.* (In publication.)
Penrose, 1946. *Trans. Faraday Soc.* (In publication.)
Whiffen and Thompson, 1946. *Trans. Faraday Soc.* (In publication.)
Horner, Taylor, Dunsmuir, Lamb and Jackson, 1946. *J. Inst. Elect. Engrs.*, **93**, Part III, 53.

THE BEHAVIOUR OF WATER UNDER HYDROSTATIC TENSION: III

By H. N. V. TEMPERLEY,
King's College, Cambridge

MS. received 9 *August* 1946

ABSTRACT. Some further experiments on the behaviour of water under tension are described, which appear to confirm the conclusions of the first two papers in this series. It appears that, under favourable conditions, water in a glass tube can support tensions as high as 60 atmospheres. A simple theoretical investigation shows that the commonly held view that the tensile strength of a liquid should be numerically equal to the "intrinsic pressure" is false. For water, a reasonable theoretical value would be 500 to 1000 atmospheres for the tensile strength, which is higher than anything that has been actually measured, though two possible explanations of this discrepancy can be suggested. In any case, it is clear that the discrepancy is much less serious than is usually supposed.

The same theoretical considerations can be applied to the experiments of Kenrick, Gilbert and Wismer on the superheating of liquids, and their results can be accounted for without any extra assumptions.

---

§1. INTRODUCTION

In two previous papers, Temperley and Chambers (1946 a), referred to as Part I, and Temperley (1946 b), referred to as Part II, some measurements of the critical tensile strength of water were described. The Berthelot method of heating water in a sealed tube until the tube filled, and then cooling it until failure occurred, was examined in detail, and evidence was obtained that quite high pressures might be developed in the tube before the final disappearance of the gas bubble by solution in the water. The assumption of zero pressure at this instant thus leads to spuriously high readings for the tension in the liquid when failure occurs. In Part I it was suggested that the high pressures were necessary to force the water to enter fissures in the glass, but in Part II clear evidence was obtained that a small gas bubble can exist for long periods in the presence of water at high pressure. A simple calculation showed that the observed rates of diffusion of gases in liquids are so low that a gas bubble would have to be compressed to a diameter of a few tenths of a millimetre to make it dissolve in a reasonable time, so that one can account for the observed high pressures in this manner also. In an attempt to decide between these two hypotheses, some experiments were carried out with a wetting agent added to the water, the original idea of which was to facilitate the flow of water into fissures in the glass. A second effect of the wetting agent was to prevent the very small bubbles that are formed at the instant of failure from joining up again immediately, as they do in plain water. The experiments described below appear to provide decisive evidence that the high pressures are necessitated by the slowness of diffusion.

An attempt, only partially successful, was made to reproduce the mechanism of the ejection of spores from ferns (King, 1944). The mechanism is the diffusion of water vapour through the cell wall, with a consequent shrinkage of the cell and the setting up of tension in the remaining liquid, the cell springing back to its normal shape when the liquid breaks. Use was made of the observation of Reekie and Aird (1945) that a tube filled with wet jeweller's rouge is permeable to water vapour but not to air.

The paper ends with a theoretical discussion of the problem of the strength of liquids. Quite simple considerations show that the assumption made by almost all writers (including the author in Parts I and II), that the so-called "intrinsic pressure" of a liquid is the same thing as its theoretical tensile strength, is false. The latter quantity is very much the smaller, though still larger than any tension that has actually been measured in water. There is a little evidence that it may be possible to approach the theoretical value by pre-compression of the water to destroy possible nuclei. The same considerations account nicely for the results obtained by Kenrick, Gilbert and Wismer (1924) on the superheating of liquids.

§2. EXPERIMENTS WITH WATER CONTAINING A WETTING AGENT

The wetting agent used was sulphonated lauryl alcohol, and I am extremely grateful to Dr. A. E. Alexander for it. It was used in approximately 1% solution, and a rough measurement (by the capillary rise method) gave a value of 25 dynes/cm. for the surface tension. Some Berthelot tubes were made up, filled with this solution and the tensile strength measured by the two methods described in Part II (the dilatometer and photo-elastic methods). Again there was satisfactory agreement between the two methods within the very limited accuracy of the photo-elastic method, but no appreciable difference between these results and those for pure water could be found, a total of twelve experiments on three different tubes giving results ranging between 20 and 60 atmospheres true tension. Mean 32 atmospheres. (A correction was made for the calorimetric effect as before.) However, a very considerable difference in the behaviour of the gas bubbles during the heating of the tube was observed, depending on whether they were large or small. At the moment of failure a multitude of very small bubbles, radii about 0·1 mm., appeared. If the tube was now immediately reheated, these tiny bubbles dissolved almost immediately, and it was found that the process could be further expedited by inverting the tube a few times, thus bringing the bubbles constantly into contact with fresh water. If, however, the tube was allowed to stand for an hour or more, the small bubbles joined up to form large ones, and the behaviour of the tube was then indistinguishable from that of one containing plain water.

The following example is typical: the filling temperature of one tube was determined with great care by the method described in Part II, the bath being heated very cautiously so that the process of dissolving the gas bubble occupied about an hour. This temperature was found to be 52° c. and the breaking temperature was determined to be $34 \pm \frac{1}{2}$° c. If, however, the tube was reheated immediately after failure of the liquid, while the bubbles were still small, it was found possible to refill it in a few minutes, even though the temperature of the bath was no more than 42° c. The exact pressure in the tube at this temperature is uncertain, but the absence of any photo-elastic effect shows that it is less than 10 atmospheres, whereas it seems to require not less than 50 atmospheres to cause a single large bubble to dissolve in a reasonable time. The breaking temperature at which the water failed under tension was the same, whichever method of filling was used. It therefore seems certain that the slowness of diffusion is responsible for the high pressures that occur during the filling of Berthelot tubes, and also that any effect of pressures, of the order of 50 to 100 atmospheres, forcing water into fissures in the glass has little effect on the tension that can subsequently be developed in the tube. Vincent's (1943) experiments on mineral oil also show that the pre-application of pressure of 100 atmospheres has little effect.

§ 3. AN ATTEMPT TO IMITATE THE MECHANISM OF SPORE-DISCHARGE

A tube was sealed at one end and a constriction formed near the other end. The tube was filled with water that had been well boiled to remove as much air as possible, and about 1 cm. of the tube above the constriction was filled with jeweller's rouge packed as tightly as possible. The open end was then connected to a calcium-chloride tube and the latter to a filter pump. At equilibrium the pressure of water vapour near the calcium chloride would be negligibly small, so one might expect tension to be developed in the water corresponding to the pressure difference that can be supported by surface tension acting in the small channels between the particles of rouge. It was shown by Reekie and Aird (1945) that such a wet pad of rouge could withstand air pressures of at least three atmospheres, but was readily permeable to water vapour. We thus seem to have a fair imitation of the mechanism of transpiration in plants.

The results of the experiments may be stated quite shortly. It was not found possible to cause pure water to fail by tension by this method, which is hardly surprising, because a tension of 40 atmospheres corresponds to an effective channel radius of only $3{\cdot}5 \times 10^{-6}$ cm. It was, however, possible to produce failure with a slight "click" if a drop of mercury had been added to the tube. The tension required to rupture a water-mercury interface is uncertain. Some experiments with Berthelot tubes containing varying quantities of mercury as well as water seemed to show that the required tension was finite, because failure occurred with formation of bubbles and an audible click, exactly as in the experiments described in Part I, when pieces of steel were introduced into the tubes, but attempts to measure this tension were not successful, which probably means that it was less than 10 atmospheres.

§ 4. THEORETICAL INVESTIGATION OF THE TENSILE STRENGTH OF LIQUIDS

[illegible] hypothesis

We consider a liquid that obeys the van der Waals equation of state. It is known (see Mayer and Mayer, *Statistical Mechanics*) that almost any law of force between the molecules leads to the van der Waals equation as a first approximation. It would not be difficult to modify the ensuing theory if a more accurate equation of state were known for a particular liquid. We write the equation of state in the form

$$(P + a/V^2)(V - b) = \boldsymbol{R}T. \qquad \ldots\ldots(1)$$

It is well known that this equation predicts a discontinuous change of state if there is any region for which $(\partial P/\partial V)_T$ is positive, such as the portion AB in figure 1, according to equation (1), for any point on this portion of the curve the thermodynamic potential takes a stationary value; but it is a maximum, not a minimum, and this portion of the curve is thus not realized in practice. The pressure at which the change of state from liquid to vapour in equilibrium with it takes place is determined by the well known rule of equal areas (dotted lines in figure 1). This rule is a simple consequence of the fact that, at equilibrium, the thermodynamic potentials of liquid and vapour must be equal, and of the thermodynamic relation $(\partial G/\partial P)_T = V$. Below this equilibrium vapour pressure, the liquid is still capable of existing in a metastable state, provided that $(\partial P/\partial V)_T$ is still negative according to equation (1). The thermodynamic potential will still be a minimum with respect to small alterations of the volume, but it will no longer be an *absolute* minimum. It follows that the limit of the metastable region is determined by the condition that $(\partial P/\partial V)_T$ vanishes, e.g. points such as A, C, and O in figure 1, because then the thermodynamic potential no longer assumes even a local minimum, but has instead a point of inflexion, so that a transition to the vapour state will certainly be able to occur without even a temporary increase in the thermodynamic potential during the process, metastable state being impossible any longer.

Thus, by differentiating equation (1) with respect to V, we find that the absolute limit of the metastable region is given by the condition

$$P - a/V^2 + 2ab/V^3 = 0. \qquad \ldots\ldots(2)$$

The corresponding pressure will be positive or negative according as V is greater or less than $2b$. At the critical temperature $(\partial^2 P/\partial V^2)_T$ vanishes as well as $(\partial P/\partial V)_T$ and we derive the well known relations

$$V_c = 3b, \qquad P_c = a/27b^2, \qquad RT_c = 8a/27b. \qquad \ldots\ldots(3)$$

If the liquid is just capable of existing in the metastable state at zero pressure (point C in figure 1), equation (2) shows that we must have $V = 2b$, and substitution in equation (1) gives us $\boldsymbol{R}T_m = a/4b$ or

$$T_m = 27T_c/32 \qquad \ldots\ldots(4)$$

for the corresponding temperature. Below this temperature, the region of metastability should extend to negative pressures, and the liquid should accordingly be capable of standing tension (point D in figure 1).

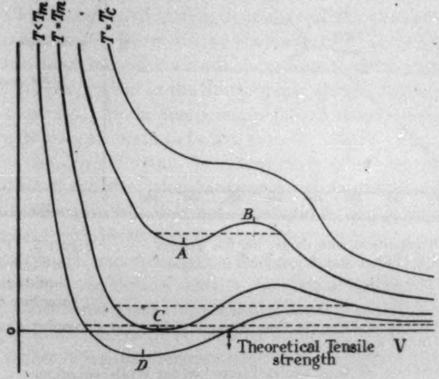


Figure 1. Isothermals of a van de Waals gas.
Dotted lines indicate equilibrium vapour pressure.

Equation (4) predicts theoretically the highest temperature at which the liquid state can exist at all at zero pressure. The superheating of a number of liquids at atmospheric pressure was studied by Kenrick, Gilbert and Wismer (1924), who found that there was a very definite limiting temperature for each liquid, at or below which it always exploded. Their values for these temperatures are plotted in figure 2, and it will be seen that there is good agreement with equation (4), even with the rather crude approximation implied by the van der Waals equation.

Equation (2) enables us to make a theoretical estimate of the behaviour of the tensile strength of a liquid as a function of temperature. The tensile strength is clearly zero at $T = T_m$ where $V = 2b$. At absolute zero, supposing the equations were still valid, we should have $V = b$, and equation (2) then gives $P = -a/b^2$, that is, the tensile strength would be 27 times the critical pressure. (Only near the absolute zero is the tensile strength comparable with the "intrinsic pressure" a/V^2.) By giving V various values between $2b$ and b, we can, by using equations (1) and (2), determine the tensile strength as a function of temperature. Because of the law of corresponding states, the theoretical curve is the same for all liquids except for scaling factors. It is plotted in figure 3 in non-dimensional form, though the positions on the curve that would be occupied by the boiling and freezing points of water have been indicated. For water, we have $T_m = 270°$ c.

It will be seen that the theoretical tensile strength of liquid water may be expected to vary between 0·1 and 0·2 a/b^2, that is, between 500 and 1000 atmospheres if we take the value of a/b^2 deduced from the critical pressure. There is no actual numerical evidence of tensions as high as this, though it is clear that the discrepancy between theory and experiment is very much less than is usually supposed. Moreover, evidence that the tensile strength of water can be increased by pre-compression or strong centrifuging has already been obtained by Harvey and others (1944.) (In Part I it was noted that no method of completely freeing water from dissolved gases had been found.) The point that appears to have been overlooked is that, at any finite temperature, the tension necessary to cause failure in a liquid may be numerically very much less than the "intrinsic pressure" a/V^2. It should be mentioned that there is an alternative method of estimating the constant a, using the latent of heat of evaporation. If the van der Waals equation held for all volumes, the work done in increasing the volume from V to infinity would be

$$RT \log(V-b) + a/V,$$

the term a/V representing the work done against the attraction of the molecules. If we identify this with the latent heat of evaporation, we get a value of

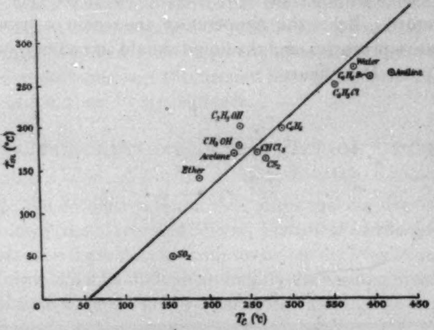


Figure 2. Comparison with theory of the results of Kenrick, Gilbert and Wismer on the superheating of liquids.

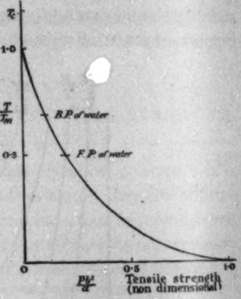


Figure 3. Theoretical tensile strength of van der Waals' liquid as a function of absolute temperature.

T_m = Limiting temperature for exsitence of liquid at atmospheric pressure.
T_c = Initial temperature.
Straight line $\frac{T_m}{T_c} = \frac{27}{32}$ deduced from van der Waals' equation.

25 000 atmospheres for the intrinsic pressure and correspondingly larger values for the tensile strength. However, this argument involves a long extrapolation, and it is known that water does not obey van der Waals' equation very well. The former argument involves the assumption that the constants a and b are independent of temperature so that they may be calculated from the critical data, and is thus also not free from objection. The theoretical tensile strength is sensitive to the value of b, as well as to that of a, but it is interesting that one can account for the experimental facts on superheating of liquids *without* assuming that these quantities vary with temperature, in fact, relation (4) is not much affected by such variations. It seems worth while to ask two further questions. First, what is the effect on the theoretical predictions of assuming the quantities a and b to be dependent on temperature; secondly, what can be said theoretically about the probable behaviour of a and b as functions of temperature?

§ 5. POSSIBILITY OF VARIATIONS OF THE VAN DER WAALS CONSTANTS WITH TEMPERATURE

The discussion of the effect of possible variations of b with temperature is quite simple. It can easily be shown from equations (1) and (2) that if b increases above the value $V_c/3$, the tensile strength at a given temperature will be smaller than if b remains constant. Now, since b is a measure of the " excluded volume " it is fairly obvious that b must decrease as the temperature rises and the collisions become more energetic, except in the limiting case of rigid molecules for which it should remain constant. As the temperature falls, b should increase, so that we might expect the tensile strength to be less than the value predicted by equations (1) and (2) from the critical data. It is also fairly clear that an increase of a, which is a measure of the attraction between the molecules, should increase the predicted tensile strength, but it is not so simple to decide whether a should increase or decrease with increasing temperature.

The following rough argument seems to show that either type of behaviour is possible. Let us suppose that we specify the state of the liquid by means of a variable r, which measures the average distance apart of neighbouring molecules. As a rough approximation, the free volume occupied by N molecules may be taken as KNr^3, where K is some constant depending on the shape of the molecules and the average number of nearest neighbours. Thus there will be a contribution to the entropy of the order of magnitude $\boldsymbol{R}\log(KNr^3)$, corresponding to the term $\boldsymbol{R}\log V$ in a perfect gas, arising from the various possible ways of arranging the molecules. Now suppose we assume that the mutual potential energy of two molecules can be expressed by means of an attraction and a repulsion, both of power-law type. The partition function will then be given by an expression of the type, qua function of r:

$$\log f = (\text{terms indep. of } r) + N\log(KNr^3) - \left(\frac{-A}{r^m} + \frac{B}{r^n}\right)\Big/ \boldsymbol{k}T. \qquad \ldots\ldots(4)$$

We are assuming that the external pressure is zero, and also that the parts of the partition function depending on the internal degrees of freedom of the molecule are not affected appreciably by changes of r. For a small external pressure, the state of absolute equilibrium is the vapour phase (r very large), but a state of metastable equilibrium is possible for small r provided that $\partial(\log f)/\partial r$ vanishes, and provided also that $\partial^2(\log f)/\partial r^2$ is negative, so that the partition function is a maximum, not a minimum. We recall that a maximum of the partition function corresponds to a minimum of the thermodynamic potential. The temperature T_m, corresponding to the limit of metastability being at zero pressure, will correspond to the vanishing of both these quantities, so that we have

$$3N\boldsymbol{k}T_m = \frac{mA}{r_m^m} - \frac{nB}{r_m^n} = \frac{m(m+1)A}{r_m^m} - \frac{n(n+1)B}{r_m^n}. \qquad \ldots\ldots(5)$$

Eliminating T_m, we find

$$r_m = \left(\frac{n^2B}{m^2A}\right)^{\frac{1}{n-m}}. \qquad \ldots\ldots(6)$$

The average attraction between molecules is proportional to the derivative of the potential energy with respect to r, and this will be a decreasing or increasing function of r according to the sign of the second derivative. The value of r for which the average attraction is a stationary function of r is given by

$$r_1 = \left(\frac{n(n+1)B}{m(m+1)A}\right)^{\frac{1}{n-m}}. \qquad \ldots\ldots(7)$$

If there were no thermal agitation, the value of r would be given by

$$r_0 = \left(\frac{nB}{mA}\right)^{\frac{1}{n-m}}. \qquad \ldots\ldots(8)$$

If we give m and n reasonable values, say 6 and 12 respectively, we see that r_1 is considerably greater than r_0, but only about 1% less than r_m. Since the volume is approximately proportional to r^3, it follows that r_1 corresponds to a temperature of about 50°–100°C. below T_m if we assume a reasonable value of 5 to 10×10^{-4} for the coefficient of expansion of the liquid per ° C. Thus, it seems likely that practically all the experiments that have been done to produce tension in liquids have taken place at a temperature below that corresponding to r_1, and thus in a region where we should expect the *average* attraction between molecules, and thus also the van der Waals constant a, to be increasing functions of the temperature. Incidentally, we also seem to have obtained an explanation of the observed fact that the van der Waals equation can be made to give fair agreement with experiment over quite considerable ranges of temperatures, because b probably changes fairly slowly with temperature, and a seems to pass through a maximum at a temperature in the liquid range.

It is hardly necessary to point out that an increase of a corresponds to an increase in tensile strength of the liquid, and this can be verified by inspection of equations (1) and (2). It is, however, not possible to say whether a liquid at a temperature of, say, 30° C. should have an effective value of a greater or less than that at its critical temperature, so that we are unable to predict the nature of the departure from the curve in figure 3 due to the variation of a with temperature. The argument indicates that the modified curve would lie above the curve in figure 3 at the temperature corresponding to r_1, and that it would cross it at some lower temperature.

It thus seems likely that any "softness" of the molecules would reduce the tensile strength below the value calculated from the critical data, but it is not possible to make any numerical prediction about the variation of the constant a with temperature, though the effect of such a variation can be calculated immediately. On the other hand, one can be fairly confident about the sign of any variation of b, and of the effect of this. It seems just worth mentioning an interpretation of this effect of "softness" of molecules. An increase in "softness" increases the volume within which the centre of a molecule may be found, and thus favours the possible formation of "holes" in a liquid. In fact, an attempt has been made by Fürth (1941) to estimate the probability of the appearance of "holes" of various sizes. The theory indeed predicts a metastable state of a liquid under tension, but the results are unlikely to be of quantitative value, because the author uses, as he himself points out, concepts such as surface energy and hydrodynamic virtual mass, which are essentially macroscopic, in connection with cavities of molecular dimensions. In fact, the theory predicts tensile strengths which are much too high.

Actual calculations have shown that the theoretical values of the temperature T_m (figure 2), are much less affected by variations of a and b with temperature than is the theoretical tensile strength (figure 3). For example, the rather extreme assumption made in Berthelot's equation of state ($a \propto 1/T$) leads to an approximately ten-fold increase in the theoretical tensile strength of water compared with the assumption that a is constant, while the ratio between T_m and T_c is changed only from 27/32 to $\sqrt{27/32}$.

§ 6. CONCLUSIONS

The conclusion reached in Part II that slowness of diffusion is the factor that causes high pressures to occur in Berthelot tubes during the filling process has been confirmed, and evidence has also been obtained that the prior application of pressures of the order of 100 atmospheres has little effect on the tension that can subsequently be developed in the tube. Theoretical considerations, supported by the experimental work on the superheating of liquids, indicate that the discrepancy between calculated and observed strengths of liquids is not nearly as great as is usually supposed. Possible explanations of the remaining discrepancy are:

(*a*) The effect of dissolved gases and other nuclei.

(*b*) The departure of liquids from an equation of state of the van der Waals type.

ACKNOWLEDGMENTS

I wish to thank Sir Lawrence Bragg for the facilities for this investigation and for helpful discussion and advice, and also Sir Geoffrey Taylor for inspiring this investigation. I should also like to thank Dr. A. E. Alexander for supplying me with the wetting agent and for helpful discussion. Finally, I wish to thank the Leverhulme Trustees for a generous research grant.

REFERENCES

FÜRTH, 1941. "On the theory of the liquid state.—II." *Proc. Camb. Phil. Soc.*, **37**, 276.

HARVEY, *et al*, 1944. "Bubble formation in animals." *J. Cellular and Comparative Physiol.*, **1**, 23.

KENRICK, GILBERT and WISMER, 1924. "The superheating of liquids." *J. Phys. Chem.*, **28**, 1297.

KING, 1944. "The spore discharge mechanism of common ferns." *Proc. Nat. Acad. Sci., Wash.*, **30**, 155.

REEKIE and AIRD, 1945. "Flow of water through very narrow channels and attempts to measure thermo-mechanical effects in water." *Nature, Lond.*, **156**, 367.

TEMPERLEY and CHAMBERS, 1946 a. "The behaviour of water under hydrostatic tension.—I." *Proc. Phys. Soc.*, **58**, 420.

TEMPERLEY, 1946 b. "The behaviour of water under hydrostatic tension.—II." *Proc. Phys. Soc.*, **58**, 436.

VINCENT, 1943. "Examination of the Berthelot method of measuring tension in liquids." *Proc. Phys. Soc.*, **55**, 376.

AN ANALYSIS OF THE CONDITIONS FOR RUPTURE DUE TO GRIFFITH CRACKS

BY H. A. ELLIOTT,
Bristol

Communicated by Professor N. F. Mott, F.R.S. ; *MS. received* 9 *September* 1946

ABSTRACT. The solutions for the problem of an infinite isotropic elastic solid stressed under tension T_0 and containing a single internal crack of length c on the plane $z=0$ are given in a form suitable for the computation of the stresses and displacements at all points. These are used to find the stress distribution on, and the displacements of, the plane situated $\frac{1}{2}a$ from the plane containing the crack. The normal stress σ_z on $z=\frac{1}{2}a$ (as found above) is plotted as a function $f(2u_z)$ of the normal displacement u_z, and τ_{rz} is small compared with σ_z.

A model is used in which the crack is considered to be bounded by the atoms centred on the planes $z=\pm\frac{1}{2}a$, these planes being the boundaries of two semi-infinite elastic solids. Equilibrium is maintained by postulating that an attractive force, $f(z)$, acts between the atoms of these bounding planes when they are $z+a$ apart. It is found that $f(z)$ approximates to the law of force expected from atomic considerations, and the condition for unstable equilibrium of the crack, i.e. a value T_0^c of T_0 such that for $T_0<T_0^c$ the crack closes (c decreases), and for $T_0>T_0^c$ the crack spreads (c increases), is found. The surface energy is calculated from the results and the equilibrium condition is found in a form similar to that of Griffith. Agreement is found with the experimental results of Griffith.

In the absence of the tension T_0, the crack cannot be maintained without an inclusion to prevent closing. Possible physical models are discussed.

§1. INTRODUCTION

THE object of this paper is to extend Griffith's theory (1921) of rupture through a more detailed consideration of the interatomic forces which resist the spread of a crack.

Griffith's theory may be expressed in the following form : suppose that a solid of Young's modulus E is subjected to a stress T_0, and that it contains a crack of length $2c$. Then the elastic energy (or that part of it which depends on c) is, per unit length,

$$-\frac{T_0^2c^2}{E}\times\pi(1-\sigma^2).$$

Griffith supposes that the total contribution to the energy which depends on c is obtained by adding the surface energy $4Sc$, where S is the surface energy per unit area. Thus we have for the energy

$$-\frac{T_0^2c^2}{E}\times\pi(1-\sigma^2)+4Sc.$$

Plotted against c, this gives a curve with a maximum at a critical value c_0; Griffith assumes that the crack will spread if c lies to the right of this maximum.

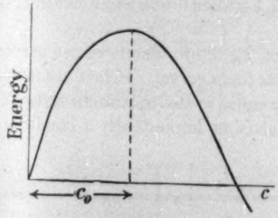


Figure 1.

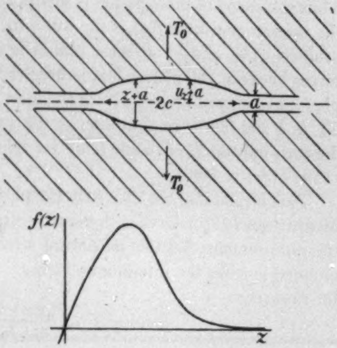


Figure 2.

In the Griffith theory it is implicitly assumed that once a crack is formed it is incapable of closing again; otherwise any crack of length less than the critical value c_0 would close. It is possible that in glass, after the silicon–oxygen bonds are broken, a displacement of the atoms in the amorphous material does in fact prevent them from reforming.

In this paper, however, we consider the fracture of a crystalline substance (e.g. a metal) along a cleavage plane. The model we take (§4) is of two semi-infinite blocks of solid attracting each other with a force $f(z)$ per unit area when they are at a distance $z+a$ from each other; $f(z)$ represents the force of attraction between two planes of atoms. $f(z)$ must be of the form $f(z)=Ez$ if z is small and $f(z)\to 0$ for $z\gg a$; here a denotes the interatomic distance.

The quantity S defined by $\int_0^\infty f(z)\,dz=2S$ represents the free energy of the surface per unit area.

We then consider two forms of crack in the boundary joining the two blocks: a penny-shaped crack of diameter $2c$, and a crack of width $2c$ but extending in the other direction to the boundaries of the solid, are treated. Such cracks will *either* close up *or* extend under an applied stress T_0, except for one value of c which gives unstable equilibrium. In any physical application of the theory to a material weakened by Griffith cracks we must consider some agency which holds the crack apart—e.g. adsorbed gas (possible models are discussed in § 5). In this paper, however, just as in the Griffith theory, we shall work out the condition for unstable equilibrium, i.e. the condition that the crack is about to extend (or contract). The method of approach is as follows: we first find the stress round a crack in a stressed elastic solid on the assumption that no stresses act across the surface of this mathematical crack, as in the work of Inglis (1911) and Griffith. We have actually used methods due to Sneddon (1947) and Elliott (1947) (§§ 2 and 3). We then calculate σ_z, the z component of stress, and τ_{rz} (or τ_{xz}), the shear stress, in the two planes which before stressing were distant $\pm \frac{1}{2}a$ from the plane of the crack. It is found that τ_{rz} (or τ_{xz}) is small compared with σ_z. Thus the shape of these planes is identical with that of the boundary of a crack, if the interatomic force $f(z)$ is the same function of z as σ_z is of $2u_z$ in the elastic problem (§ 4), and σ_z on these planes is now assumed to be due to interatomic forces between the planes. u_z is here the displacement in the z direction of the planes mentioned above. Near the apex of the crack, σ_z and u_z depend only on T_0, E, σ and $(c/a)^{\frac{1}{2}}$ (to the order of the approximation used), so the (σ_z, u_z) curve depends only on T_0, E, σ and $(c/a)^{\frac{1}{2}}$ for values of u_z between $0{\cdot}05a$ and a (when C is of the order of 10^4a).

This means that we have only the parameters, T_0, $(c/a)^{\frac{1}{2}}$, which we can use to identify our (σ_z, u_z) curve with the true interatomic force curve. In fact, we make the requirement that the maximum stress σ_z is equal to the maximum stress P obtainable from the interatomic forces. This gives us immediately a condition for rupture:—

$$T_0\sqrt{c} = \frac{4\pi}{7}P\sqrt{a} \quad (\sigma = 0{\cdot}25)$$

for the case of a penny-shaped crack, or ......(1.1, 1)

$$T_0\sqrt{c} = \frac{8}{7}P\sqrt{a} \quad (\sigma = 0{\cdot}25)$$

for the case of a Griffith crack.

A check on the validity of this identification is given by the fact that we have now determined the surface energy S in terms of P, E and a. In fact we find,

$$\frac{S\,.\,E}{P^2\,.\,a} \approx 1 \qquad (\sigma = 0{\cdot}25). \qquad \text{......(1.1, 2)}$$

Approximate agreement is found in the case of the experiments of Griffith (1924).

This relationship permits the expression of our rupture condition in the same form as that of Griffith, i.e. for the penny-shaped crack and the Griffith crack respectively:

$$\left.\begin{aligned} T_0\sqrt{c} &= \frac{4\pi}{7\sqrt{2}}\sqrt{(E\,.\,S)} = 1{\cdot}27\sqrt{(E\,.\,S)} \quad (\text{for } \sigma = 0{\cdot}25) \\ \text{or} \quad T_0\sqrt{c} &= \frac{8}{7\sqrt{2}}\sqrt{(E\,.\,S)} = 0{\cdot}81\sqrt{(E\,.\,S)} \end{aligned}\right\} \qquad \text{......(1.1, 3)}$$

as compared with Griffith's result:

$$T_0\sqrt{c} = \sqrt{\left(\frac{2E\,.\,S}{\pi(1-\sigma^2)}\right)} \qquad \ldots\ldots(1.1,4)$$

$$= 0{\cdot}824\sqrt{(E\,.\,S)} \quad (\text{for } \sigma = 0{\cdot}25). \qquad \ldots\ldots(1.1,5)$$

§ 2. ANALYSIS OF STRESS AND STRAIN FOR A PENNY-SHAPED CRACK

2.1. In the three dimensional case, we consider a penny-shaped crack, bounded by the circle $x^2+y^2=c^2$ in the plane $z=0$. Hence the distribution of stress and displacement is the same as that for the semi-infinite elastic medium bounded by the plane $z=0$, the boundary conditions being for a crack maintained by internal pressure p_0:—

$$\left.\begin{array}{ll} \text{i.} & \tau_{xy}=\tau_{yz}=\tau_{rz}=0 \text{ on } z=0 \text{ for all } x, y; \\ \text{ii.} & \sigma_z=-p_0 \text{ on } z=0, \quad r=\sqrt{(x^2+y^2)}\leqslant c \\ & u_z=0 \text{ on } z=0, \quad r\geqslant c. \end{array}\right\} \qquad \ldots\ldots(2.1,1)$$

We now employ cylindrical polar co-ordinates (r, θ, z) and we use the solutions in the form given by Sneddon (1947) (equations 3.6, 2–3.6, 9):

$$u_r = +\frac{2p_0c(1+\sigma)(1-2\sigma)}{\pi E}\int_0^\infty \left(1-\frac{\zeta\eta}{1-2\sigma}\right)\frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right)e^{-\zeta\eta}J_1(\rho\eta)\,d\eta$$

$$= +\frac{2p_0c(1+\sigma)(1-2\sigma)}{\pi E}\left[\frac{2(1-\sigma)\zeta}{1-2\sigma}S_0{}^1-\frac{\rho}{2}S_0{}^0+\frac{\rho}{2}S_0{}^2-\frac{\zeta}{1-2\sigma}c_1{}^1\right], \qquad \ldots\ldots(2.1,2)$$

$$u_z = -\frac{4p_0c(1-\sigma^2)}{\pi E}\int_0^\infty \left[1+\frac{\zeta\eta}{2(1-\sigma)}\right]\frac{d}{d\eta}\left(\frac{\sin\eta}{\eta}\right)e^{-\zeta\eta}J_0(\rho\eta)\,d\eta$$

$$= \frac{4p_0c(1-\sigma^2)}{\pi E}\left[1-\frac{\zeta^2}{2(1-\sigma)}S_1{}^0-\frac{\zeta\rho}{2(1-\sigma)}S_1{}^1-\zeta\frac{1-2\sigma}{2(1-\sigma)}S_0{}^0-\rho S_0{}^1\right], \qquad \ldots\ldots(2.1,3)$$

$$\sigma_z = \frac{2p_0}{\pi}[C_1{}^0-S_0{}^0+\zeta C_2{}^0-\zeta S_1{}^0], \qquad \ldots\ldots(2.1,4)$$

$$\tau_{rz} = \frac{2p_0}{\pi}[C_2{}^1-S_1{}^1], \qquad \ldots\ldots(2.1,5)$$

$$\sigma_r+\sigma_\theta+\sigma_z = \frac{4(1+\sigma)p_0}{\pi}[C_1{}^0-S_0{}^0], \qquad \ldots\ldots(2.1,6)$$

$$\sigma_\theta-\sigma_r = \frac{2p_0}{\pi}[(1-2\sigma)(C_1{}^2-S_1{}^2)-\zeta(C_2{}^2-S_1{}^2)], \qquad \ldots\ldots(2.1,7)$$

where

$$\zeta = \frac{z}{c}, \qquad \rho = \frac{r}{c}$$

and

$$\left.\begin{array}{l} Z_n^m = C_n^m(\rho,\zeta)-iS_n^m(\rho,\zeta) \\ \quad = \displaystyle\int_0^\infty \eta^{n-1}e^{-\zeta\eta}J_m(\rho\eta)\,d\eta. \end{array}\right\} \qquad \ldots\ldots(2.1,8)$$

If we now write

$$r=1+\zeta^2, \qquad \zeta\tan\theta=1, \qquad \ldots\ldots(2.1,9)$$

$$R^2=(\rho^2+\zeta^2-1)^2+4\zeta^2, \qquad 2\zeta\cot\phi=\rho^2+\zeta^2-1,$$

we have

$$\left.\begin{aligned} C_1^0&=R^{-\frac{1}{2}}\cos\tfrac{1}{2}\phi,\\ C_2^0&=rR^{-\frac{3}{2}}\cos(\tfrac{3}{2}\phi-\theta),\\ C_2^1&=\rho R^{-\frac{3}{2}}\cos\tfrac{3}{2}\phi,\\ S_0^0&=\tan^{-1}\frac{R^{\frac{1}{2}}\sin\frac{1}{2}\phi+r\sin\theta}{R^{\frac{1}{2}}\cos\frac{1}{2}\phi+r\cos\theta},\\ S_1^0&=R^{-\frac{1}{2}}\sin\tfrac{1}{2}\phi,\\ S_0^1&=\frac{1}{\rho}(1-R^{\frac{1}{2}}\sin\tfrac{1}{2}\phi),\\ \text{and}\quad S_1^1&=\frac{r}{\rho}R^{-\frac{1}{2}}\sin(\theta-\tfrac{1}{2}\phi). \end{aligned}\right\} \qquad \ldots\ldots(2.1,10)$$

2.2. The equations given in §2.1 give the complete system of stress and displacement at all points in the medium for a crack maintained by an internal pressure p_0 in a body free from stress at infinity. For the case of Griffith's fracture we are concerned with a body under surface forces, the faces of any crack being assumed free surfaces in the elastic-theory solutions.

If these surface forces are given by tensions (or pressures), the solutions are easily found from those in 2.1. Thus for $T_z=T_0$, $T_r=T_1$ we have

$$\begin{aligned} u_r=\frac{T_0}{E}\bigg\{\frac{2(1+\sigma)(1-2\sigma)c}{\pi E}\bigg[&\frac{2(1-\sigma)\zeta}{1-2\sigma}S_0^1-\frac{\rho}{2}S_0^0\\ &+\frac{\rho}{2}S_0^2-\frac{\zeta}{1-2\sigma}C_1^1\bigg]-\sigma r\bigg\}+\frac{T_1 r}{E}, \end{aligned} \qquad \ldots\ldots(2.2,1)$$

$$\begin{aligned} u_z=\frac{T_0}{E}\bigg\{z+\frac{4(1-\sigma^2)c}{\pi}\bigg[&1-\frac{\zeta^2}{2(1-\sigma)}S_1^0-\frac{\zeta\rho}{2(1-\sigma)}S_1^1\\ &-\zeta\frac{1-2\sigma}{2(1-\sigma)}S_0^0-\rho S_0^1\bigg]\bigg\}-T_1\frac{\sigma z}{E}, \end{aligned} \qquad \ldots\ldots(2.2,2)$$

$$\sigma_z=T_0\left\{1+\frac{2}{\pi}[C_1^0-S_0^0+\zeta C_2^0-\zeta S_1^0]\right\}, \qquad \ldots\ldots(2.2,3)$$

$$\tau_{rz}=T_0\frac{2}{\pi}\zeta[C_2^1-S_1^1], \qquad \ldots\ldots(2.2,4)$$

$$\sigma_r+\sigma_\theta+\sigma_z=T_0\left\{1+\frac{4(1+\sigma)}{\pi}(C_1^0-S_0^0)\right\}+T_1, \qquad \ldots\ldots(2.2,5)$$

$$\sigma_r-\sigma_\theta=\frac{2T_0}{\pi}\{(1-2\sigma)(C_1^2-S_0^0)-\zeta(C_2^2-S_1^2)\}-T_1. \qquad \ldots\ldots(2.2,6)$$

2.3. Under these conditions the shape of the crack becomes an oblate ellipsoid

$$\frac{x^2+y^2}{c^2}+\frac{z^2}{\epsilon^2}=1, \qquad \ldots\ldots(2.3,1)$$

where

$$\epsilon=[u_z]_{r=0,z=0}=\frac{4(1-\sigma^2)T_0c}{\pi E}. \qquad \ldots\ldots(2.3,2)$$

More exactly stated, this is the shape of the boundary found by the methods of elastic theory. However, in the actual physical case we are more interested in the displacement of the first plane of atomic centres on either side of the cleavage plane, and these lie at a finite distance, $\frac{1}{2}a$ ($a=$ lattice spacing) from the theoretical boundary, $z=0$.

We know from the Griffith theory (Griffith, 1921) that in normal materials the length of crack required to give the observed fracture stress values is of the order of 10^{-5} cm. in length. This is large compared with the atomic spacing (10^{-8} cm.), so that for our calculations for the displacement of the first plane of atomic centres we can disregard powers of $(a/2c)$ above the lowest found in the expression for any stress or displacement. We then expect the approximate results so obtained to be correct to the order of one part in 10^3, provided that all the coefficients in the full expansions in terms of $(a/2c)^n$ are of an order not greater than that of the coefficient of the retained terms.

2.4. To find the displacement of the plane $z=\frac{1}{2}a$ we put $\zeta=a/2c=\zeta_0$ in the expressions of (2.1, 10). Thus we obtain to sufficiently high order in ζ_0:—

Domain 1. At great distances from the crack:

$$\left.\begin{aligned} &r\to\infty \quad \text{and so} \quad \zeta_0 \ll \rho^2-1, && S_0^{\,0}=\tan^{-1}(\rho^2-1)^{-\frac{1}{2}},\\ &C_1^{\,0}=(\rho^2-1)^{-\frac{1}{2}}, && S_1^{\,0}=\zeta_0/(\rho^2-1)^{\frac{3}{2}},\\ &C_2^{\,0}=\zeta_0^{\,2}(\rho^2-1)^{-\frac{3}{2}}, && S_0^{\,1}=\rho^{-1}\{1-\zeta_0/(\rho^2-1)^{\frac{1}{2}}\},\\ &C_2^{\,1}=\rho(\rho^2-1)^{-\frac{3}{2}}, && S_1^{\,1}=\rho^{-1}(\rho^2-1)^{-\frac{1}{2}}. \end{aligned}\right\} \qquad \text{......(2.4, 1)}$$

Domain 2. Near the centre of the crack:

$$\left.\begin{aligned} &r\ll c \quad \text{and so} \quad \rho^2\ll 1, && S_0^{\,0}=\tan^{-1}[\zeta_0(1-\rho^2)^{\frac{1}{2}}],\\ &C_1^{\,0}=\zeta_0(1-\rho^2)^{-\frac{3}{2}}, && S_1^{\,0}=(1-\rho^2)^{-\frac{1}{2}},\\ &C_2^{\,0}=-(1-\rho^2)^{-\frac{3}{2}}, && S_0^{\,1}=\{1-(1-\rho^2)^{\frac{1}{2}}-\tfrac{1}{2}\zeta_0^{\,2}\rho^2(1-\rho^2)^{-\frac{3}{2}}\},\\ &C_2^{\,1}=-3\zeta_0\rho(1-\rho^2)^{-\frac{5}{2}}, && S_1^{\,1}=\zeta_0\rho(1-\rho^2)^{-\frac{3}{2}}. \end{aligned}\right\} \qquad \text{......(2.4, 2)}$$

Domain 3. Near the edge of the crack:

$r\simeq c$ and so $\rho^2\simeq 1$; ρ^2-1 is of the same order as ζ_0.

Write

$$\rho-1=\mu\zeta_0. \qquad \text{......(2.4, 3)}$$

Then μ is of order unity in the domain where the following approximations hold. We may note that $\frac{1}{2}\mu$ is the number of lattice spacings between the point $(r,\frac{1}{2}a)$ or (ρ,ζ_0) being considered and the "edge" of the crack, $r=c$ ($\rho=1$), in the case of a simple cubic structure.

When

$$\left.\begin{aligned} &\rho-1=\mu\zeta_0,\\ &C_1^{\,0}=\zeta_0^{-\frac{1}{2}}2^{-1}(1+\mu^2)^{-\frac{1}{4}}[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}},\\ &C_2^{\,0}=\zeta_0^{-\frac{1}{2}}2^{-2}(1+\mu^2)^{-\frac{3}{4}}\{[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}+\mu[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\},\\ &C_2^{\,1}=\zeta_0^{-\frac{3}{2}}2^{-1}(1+\mu^2)^{-\frac{3}{4}}\{\mu[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}-[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\},\\ &S_0^{\,0}=\tan^{-1}\{2^{-1}(1+\mu^2)^{-\frac{1}{4}}[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{-\frac{1}{2}}\zeta_0^{-\frac{1}{2}}\},\\ &S_1^{\,0}=2^{-1}(1+\mu^2)^{-\frac{1}{4}}\{1-\mu(1+\mu^2)^{-\frac{1}{2}}\}^{\frac{1}{2}}\zeta_0^{-\frac{1}{2}},\\ &S_0^{\,1}=1-(1+\mu^2)^{\frac{1}{4}}\{1-\mu(1+\mu^2)^{-\frac{1}{2}}\}^{\frac{1}{2}}\zeta_0^{\frac{1}{2}}\rho,\\ &S_1^{\,1}=2^{-1}(1+\mu^2)^{-\frac{1}{4}}\{1+\mu(1+\mu^2)^{-\frac{1}{2}}\}^{\frac{1}{2}}\zeta_0^{-\frac{1}{2}}\rho. \end{aligned}\right\} \qquad \text{......(2.4, 4)}$$

Putting these values in the expressions for u_z, σ_z, τ_{rz}, we have for $z=c\zeta_0$, $\zeta_0=a/2c$

$$u_z=\frac{T_0c\zeta_0}{E}-\frac{T_1\sigma c\zeta_0}{E}+\frac{4(1-\sigma^2)cT_0}{\pi E}U_z, \qquad \ldots\ldots(2.4,5)$$

where
$$U_z=(1-\rho^2)^{\frac{1}{2}}-\frac{1-2\sigma}{2(1-\sigma)}\zeta_0\tan^{-1}[\zeta_0^{-1}(1-\rho^2)^{\frac{1}{2}}] \qquad \ldots\ldots(2.4,6)$$

if $\rho\ll 1$, i.e. near centre of crack,

$$U_z=\left\{[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}-\frac{1}{4(1-\sigma)}[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}(1+\mu^2)^{-\frac{1}{2}}\right\}\zeta^{\frac{1}{2}} \qquad \ldots\ldots(2.4,7)$$

if $\rho-1=\mu\zeta_0$, $\mu=0(1)$, i.e. at edge of crack,

$$U_z=\frac{1-2\sigma}{2(1-\sigma)}\left\{\frac{1}{(\rho^2-1)^{\frac{1}{2}}}-\tan^{-1}\frac{1}{(\rho^2-1)^{\frac{1}{2}}}\right\}\zeta_0 \qquad \ldots\ldots(2.4,8)$$

if $\rho\gg 1$ at great distances from the crack.

$$\sigma_z=T_0+\frac{2T_0}{\pi}\Sigma_z, \qquad \ldots\ldots(2.4,9)$$

where
$$\Sigma_z=\tan^{-1}[\zeta_0^{-1}(1-\rho^2)^{\frac{1}{2}}], \qquad \ldots\ldots(2.4,10)$$

if $\rho\ll 1$,

$$\Sigma_z=(1+\mu^2)^{-\frac{3}{2}}\{[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}[1+2(1+\mu^2)]+\mu[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\}\zeta_0^{-\frac{1}{2}}/4 \qquad \ldots\ldots(2.4,11)$$

if $\rho-1=\mu\zeta_0$, $\mu=0(1)$,

$$\Sigma_z=(\rho^2-1)^{-\frac{1}{2}}-\sin^{-1}(1/\rho) \qquad \ldots\ldots(2.4,12)$$

if $\rho\gg 1$, and
$$\tau_{rz}=\frac{2T_0}{\pi}T_{rz}, \qquad \ldots\ldots(2.4,13)$$

where
$$T_{rz}=-\zeta_0^2\rho(1-\rho^2)^{-\frac{3}{2}}\{1+3(1-\rho^2)^{-1}\}, \qquad \ldots\ldots(2.4,14)$$

if $\rho^2\ll 1$,

$$T_{rz}=(1+\mu^2)^{-\frac{3}{2}}\{\mu[1+\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}-[1-\mu(1+\mu^2)^{-\frac{1}{2}}]^{\frac{1}{2}}\}\zeta_0^{-\frac{1}{2}}/4, \qquad \ldots\ldots(2.4,15)$$

if $\rho-1=\mu\zeta_0$,

$$T_{rz}=\zeta_0\{\rho(\rho^2-1)^{-\frac{3}{2}}-\rho^{-1}(\rho^2-1)^{-\frac{1}{2}}\} \qquad \ldots\ldots(2.4,16)$$

if $\rho^2\gg 1$.

2.5. We are now able to compute the displacement u_z and the stresses σ_z and τ_{rz} due to any T_0, length of crack, c and atomic spacing a, i.e. any ζ_0. We are particularly interested in the region near the edge of the crack ($\rho\simeq 1$) (*Domain* 3), i.e. in the domain $-10\leqslant\mu\leqslant+10$ and especially $-1\leqslant\mu\leqslant+1$. The results for this domain are therefore given in detail (table 1, where values of A, C, D and F are given), u_z, σ_z and τ_{rz} being given by

$$u_z=\frac{T_0c\zeta_0}{E}-\frac{\sigma T_1c\zeta_0}{E}+\frac{4(1-\sigma^2)aT_0}{2\pi E}\left\{C-\frac{D}{4(1-\sigma)}\right\}\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(2.5,1)$$

$$\sigma_z=T_0+\frac{2T_0}{\pi}A\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(2.5,2)$$

$$\tau_{rz}=\frac{2T_0}{\pi}F\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(2.5,3)$$

for $\rho-1=\mu\zeta_0$, $-10\leqslant\mu\leqslant+10$.

Table 1

μ	r	A	C	D	F
−10	$c-5{\cdot}0\,a$	+0·00004	+4·4775	+0·02204	−0·00165
− 9	$c-4{\cdot}5\,a$	0·00026	4·2493	0·02602	−0·00214
− 8	$c-4{\cdot}0\,a$	0·00033	4·0076	0·03100	−0·00255
− 7	$c-3{\cdot}5\,a$	0·00048	3·7505	0·03771	−0·00400
− 6	$c-3{\cdot}0\,a$	0·00081	3·4761	0·04731	−0·00578
− 5	$c-2{\cdot}5\,a$	0·00105	3·1774	0·06171	−0·00895
− 4	$c-2{\cdot}0\,a$	0·00314	2·8495	0·08489	−0·01513
− 3	$c-1{\cdot}5\,a$	0·00802	2·4825	0·12742	−0·02918
− 2	$c-1{\cdot}0\,a$	0·02078	2·0593	0·21726	−0·06775
− 1·0	$c-0{\cdot}5\,a$	0·14716	1·5536	0·45516	−0·19422
− 0·4	$c-0{\cdot}2\,a$	0·44936	1·21556	0·76386	−0·30907
− 0·2	$c-0{\cdot}1\,a$	0·60528	1·10460	0·88784	−0·30305
− 0·1	$c-0{\cdot}05\,a$	0·68016	1·05230	0·94536	−0·28259
+ 0·0	$c-0{\cdot}00\,a$	0·75000	1·00000	1·00000	−0·25000
+ 0·1	$c+0{\cdot}05\,a$	0·80612	0·95006	1·04708	−0·20808
+ 0·2	$c+0{\cdot}1\,a$	0·84560	0·90544	1·08064	−0·16135
+ 0·4	$c+0{\cdot}2\,a$	0·87329	0·82288	1·12832	−0·06236
+ 0·6	$c+0{\cdot}3\,a$	0·85027	0·75248	1·14216	+0·02770
+ 1·0	$c+0{\cdot}5\,a$	0·74354	0·64370	1·09850	+0·07541
+ 2	$c+1{\cdot}0\,a$	0·51600	0·48582	0·92053	+0·08119
+ 3	$c+1{\cdot}5\,a$	0·42165	0·40297	0·78501	+0·05569
+ 4	$c+2{\cdot}0\,a$	0·36080	0·35003	0·69109	+0·03949
+ 5	$c+2{\cdot}5\,a$	0·31327	0·31468	0·62312	+0·02932
+ 6	$c+3{\cdot}0\,a$	0·29150	0·28770	0·57147	+0·02285
+ 7	$c+3{\cdot}5\,a$	0·26909	0·26667	0·52916	+0·01837
+ 8	$c+4{\cdot}0\,a$	0·25142	0·24996	0·49706	+0·01517
+ 9	$c+4{\cdot}5\,a$	0·23672	0·23558	0·46925	+0·01280
+10	$c+5{\cdot}0\,a$	0·22444	0·22150	0·44556	+0·01098

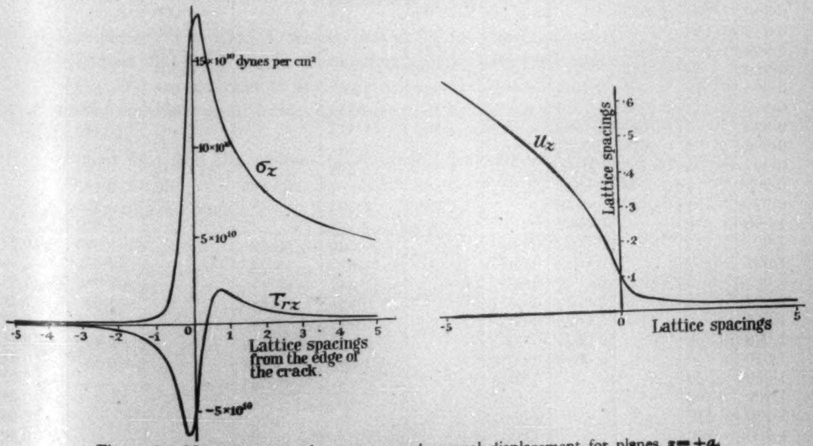


Figure 3. Normal stress, shear stress and normal displacement for planes $z=\pm a$, $T_0=10^{10}$ dynes/cm.^{2}, $E=10^{12}$ dynes/cm.^{2}, $\sigma=0{\cdot}25$. $2c=1000a$.

The displacement u_z and distribution of the stresses along the first plane of atomic centres (for $T_1=0$, $T_0=10^{10}$ dynes/sq. cm., $E=10^{12}$ dynes/cm. $\sigma=0{\cdot}25$) is shown in figure 3 where σ_z, τ_{rz} and u_z are shown for distances measured in terms of atomic radii from the edge of the crack, for a crack 1000 lattice spacings in length (see also table 2).

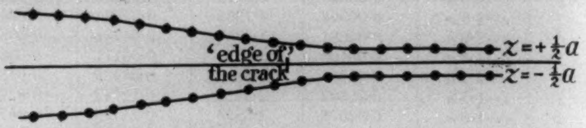


Figure 4. Configuration of atoms near the edge of the crack.

In figure 4 the displacement is shown for a crack about 2500 lattice spacings long in a glass typical of that used by Griffith (1921); this length is approximately that required to give the rupture strength observed by Griffith.

It will be seen that the change-over between the three regions is smooth and occurs about 5 lattice spacings from the crack edge. This is shown in table 2 for a crack 1000 atoms long, the values of U_z, Σ_z and T_{rz} being given as calculated from the expressions valid in each of the three regions.

Table 2

$\rho=\frac{r}{c}$	μ	U_z			Σ_z			T_{rz}		
		$\rho\ll 1$	$\rho\simeq 1$	$\rho\gg 1$	$\rho\ll 1$	$\rho\simeq 1$	$\rho\gg 1$	$\rho\ll 1$	$\rho\simeq 1$	$\rho\gg 1$
0·800	−200	0·5995			−1·569			−0 ·000		
0·850	−150	0·5263			−1·569			− 0·000		
0·900	−100	0·4354			−1·569			− 0·000		
0·950	− 50	0·3118			−1·568			− 0·001		
0·990	− 10	0·1406	0·1408		−1·564	− 1·563		− 0·054	−0·052	
0·992	− 8	0·1258	0·1259		−1·563	− 1·553		− 0·095	−0·090	
0·994	− 6	0·1089	0·1089		−1·562	− 1·536		− 0·191	−0·183	
0·996	− 4	0·0888	0·0887		−1·559	− 1·461		− 0·527	−0·480	
0·997	− 3	0·0769	0·0766		−1·558	− 1·304		− 1·079	−0·922	
0·998	− 2	0·0628	0·0620		−1·555	− 0·898		− 2·962	−2·147	
0·999	− 1	0·0442	0·0438		−1·548	+ 4·103		−16·772	−6·139	
1·000	0		0·0206			+22·179			−7·902	
1·001	+ 1		0·0083	0·0068		+21·980	+20·29		+2·384	+10·377
1·002	+ 2		0·0052	0·0047		+14·812	+13·92		+2·566	+ 3·666
1·003	+ 3		0·0040	0·0037		+11·842	+11·11		+1·760	+ 1·993
1·004	+ 4		0·0033	0·0031		+ 9·928	+ 9·43		+1·249	+ 1·293
1·006	+ 6		0·0025	0·0025		+ 7·757	+ 7·44		+0·722	+ 0·701
1·008	+ 8		0·0021	0·0021		+ 6·505	+ 6·274		+0·479	+ 0·455
1·010	+ 10		0·0018	0·0018		+ 5·667	+ 5·628		+0·347	+ 0·348
1·050	+ 50			0·0006			+ 1·863			+ 0·029
1·100	+100			0·0004			+ 1·041			+ 0·009
1·150	+150			0·0002			+ 0·707			+ 0·004
1·200	+200			0·0002			+ 0·522			+ 0·003

§ 3. ANALYSIS OF STRESS AND STRAIN FOR THE TWO-DIMENSIONAL CRACK

3.1. The problem of the crack formed by a cut from $x=-c$ to $x=+c$ in an elastic medium is easily seen to be equivalent to the two-dimensional problem of an infinite semi-plane with the boundary conditions

$$\left.\begin{aligned} \tau_{xz}&=0 && \text{on} \quad z=0, \\ \sigma_z&=-p_0 && \text{on} \quad z=0, \quad |x|<c, \\ u_z&=0 && \text{on} \quad z=0, \quad |x|\geqslant c, \end{aligned}\right\} \qquad \ldots\ldots(3.1,1)$$

if the crack is maintained by an internal pressure.

We use the solution of this problem as found by Westergaard (1939) and Elliott and Sneddon (1947). Changing the problem to that of a crack in a body in tension (T_0) as we did in the three-dimensional case, we have the solutions

$$\tfrac{1}{2}(\sigma_x+\sigma_z)=T_0\left\{\frac{r}{r_1^{\frac{1}{2}}r_2^{\frac{1}{2}}}\cos\left(\theta-\tfrac{1}{2}\theta_1-\tfrac{1}{2}\theta_2\right)\right\}, \qquad \ldots\ldots(3.1,2)$$

$$\tfrac{1}{2}(\sigma_z-\sigma_x)=T_0\frac{r\sin\theta}{c}\left(\frac{c^2}{r_1r_2}\right)^{\frac{3}{2}}\sin\tfrac{3}{2}(\theta_1+\theta_2), \qquad \ldots\ldots(3.1,3)$$

$$\tau_{xz}=T_0\frac{r\sin\theta}{c}\left(\frac{c^2}{r_1r_2}\right)^{\frac{3}{2}}\cos\tfrac{3}{2}(\theta_1+\theta_2), \qquad \ldots\ldots(3.1,4)$$

where $$x+iz=re^{i\theta}, \quad x-c+iz=r_1e^{i\theta_1}, \quad x+c+iz=r_2e^{i\theta_2}, \qquad \ldots\ldots(3.1,5)$$

and for the "shape" of the crack we have the ellipse

$$\frac{x^2}{c^2}+\left(\frac{u_z}{\epsilon}\right)^2=1, \qquad \ldots\ldots(3.1,6)$$

where $$\epsilon=2(1-\sigma^2)T_0c/E. \qquad \ldots\ldots(3.1,7)$$

By comparison with (2.3, 2) we see that u_z for the two-dimensional problem will be $\pi/2$ times that for the three-dimensional case with the same applied tensions.

3.2. We now proceed to find the stresses and displacements of the first planes of atomic centres from the cleavage plane. As before, we write $\zeta_0=a/2c$ and treat ζ_0 as small.

Let us first find the stresses in the neighbourhood of the edge of the crack. If μ is measured from the crack edge as in the three-dimensional case, i.e. $(x/c)-1=\mu\zeta_0$, we have

$$\left.\begin{aligned} r_2&\simeq 2c+\mu a/2, \\ r&\simeq c+\mu a/2, \\ r_1\cos\theta_1&\simeq x_1=\mu a, \\ r_2\cos\theta_2&\simeq z_1=a/2, \end{aligned}\right\} \qquad \ldots\ldots(3.2,1)$$

and so

$$\tfrac{1}{2}(\sigma_x+\sigma_z)=T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}}{2(1+\mu^2)^{\frac{1}{2}}}, \qquad \ldots\ldots(3.2,2)$$

$$\tfrac{1}{2}(\sigma_z-\sigma_x)=T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}+\mu[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}, \qquad \ldots\ldots(3.2,3)$$

giving $$\sigma_z = T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}[1+2(1+\mu^2)]+\mu[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}, \qquad \ldots\ldots(3.2,4)$$

$$\sigma_x = T_0\zeta_0^{-\frac{1}{2}}\frac{[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}[2(1+\mu^2)-1]-\mu[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}, \qquad \ldots\ldots(3.2,5)$$

and $$\tau_{xz} = T_0\zeta_0^{-\frac{1}{2}}\frac{\mu[(1+\mu^2)^{\frac{1}{2}}+\mu]^{\frac{1}{2}}-[(1+\mu^2)^{\frac{1}{2}}-\mu]^{\frac{1}{2}}}{4(1+\mu^2)^{\frac{3}{2}}}. \qquad \ldots\ldots(3.2,6)$$

By comparison with the results for the three-dimensional case we see that

$$\left.\begin{aligned}[\sigma_z]_3 &= \frac{2}{\pi}[\sigma_z]_2,\\ [\tau_{rz}]_3 &= \frac{2}{\pi}[\tau_{rz}]_2,\end{aligned}\right\} \qquad \ldots\ldots(3.2,7)$$

in this region, and as we have seen above (3.1'),

$$[u_z]_3 = \frac{2}{\pi}[u_z]_2.$$

To this accuracy then

$$[\sigma_z]_2 = T_0 A\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(3.2,8)$$

$$[\tau_{rz}]_2 = T_0 F\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(3.2,9)$$

$$[u_z]_2 = \frac{(1-\sigma^2)aT_0}{E}\left\{C-\frac{D}{4(1-\sigma)}\right\}\zeta_0^{-\frac{1}{2}}, \qquad \ldots\ldots(3.2,10)$$

where A, C, D, F have the same values as in the three-dimensional case. Hence from any results which we derive for the three-dimensional case with approximations of this accuracy, we can obtain the results for the two-dimensional case by using the multiplying factor $\frac{1}{2}\pi$.

§ 4. CONSTRUCTION OF A MODEL AND THE DERIVATION OF CONDITIONS FOR RUPTURE

4.1. We now use the results of the above analysis to find a suitable model for a crack in a real solid.

We may replace our infinite ideal elastic solid under tension with the theoretical crack $x^2+y^2 \leqslant c^2$, $z=0$ (or $|x| \leqslant c$, $z=0$) by two semi-infinite blocks originally bounded by the planes $z=\pm\frac{1}{2}a$ and a law of force between the two planes such that it maintains forces equal to $[\sigma_z]_{z=\pm\frac{1}{2}a}$ and $[\tau_{rz}]_{z=\pm\frac{1}{2}a}$ on the two planes. As τ_{rz} is small compared with σ_z and is along the surface, we shall neglect it; we shall suppose that, if σ_z is plotted against μ_z, and the resulting function is taken as the law of force, the strains are approximately those calculated above.

If we graph σ_z against u_z we find that for $c \simeq 10^4 a$ the form of the curve between $u_z = 0{\cdot}05a$ and a is given by

$$\left.\begin{aligned}\sigma_z &= \frac{2T_0}{\pi}A\zeta_0^{-\frac{1}{2}},\\ u_z &= \frac{4(1-\sigma^2)}{\pi E}\cdot\frac{a}{2}\cdot T_0\left\{C-\frac{D}{4(1-\sigma)}\right\}\zeta_0^{-\frac{1}{2}}\end{aligned}\right\} \qquad \ldots\ldots(4.1,1)$$

in the case of the three-dimensional crack (see figure 5), and

$$\left.\begin{aligned}\sigma_z &= T_0 A \zeta_0^{-\frac{1}{2}},\\ u_z &= \frac{2(1-\sigma^2)}{E}\cdot\frac{a}{2}\cdot T_0\left\{C-\frac{D}{4(1-\sigma)}\right\}\zeta_0^{-\frac{1}{2}}\end{aligned}\right\} \quad \ldots\ldots(4.1, 2)$$

in the two-dimensional case, i.e. by the expressions which hold near the edge of the crack, to an accuracy better than 1%.

It will be noted that the above expressions depend only on a, E, σ and $T_0\sqrt{(2c/a)}$ as parameters, A, C and D being independent of any properties of the body.

If we consider σ_z for very small values of u_z we find that $\sigma_z/(2u_z/a)$ tends to the value E. This is to be expected, as for large distances from the crack, i.e. very small u_z, the extension is merely given by Hooke's law for simple tension. For large values of u_z, $\sigma_z \sim \left[1-\frac{2}{\pi}\tan^{-1}\lambda u_z\right]$ and tends to zero as u_z tends to infinity.

We thus have a law of force, σ_z as a function of z, which behaves in the manner of the interatomic forces; the initial rate of increase of stress with strain is E, there exists a maximum attractive force, and the stress dies away to zero for infinite strain. If, in fact, the law of force we have derived is to correspond to the true interatomic forces expected, then the maximum stresses (P) must be equal. This condition determines the parameter $T_0\sqrt{(2c/a)}$, i.e. $T_0\zeta_0^{-\frac{1}{2}}$ appearing in (4.1, 1) or (4.1, 2), in fact

$$T_0\sqrt{(2c/a)} = \frac{\pi}{2A_{\max}}P \quad \text{from the three-dimensional case,} \quad \ldots\ldots(4.1, 3)$$

or

$$T_0\sqrt{(2c/a)} = \frac{1}{A_{\max}}P \quad \text{from the two-dimensional case,} \quad \ldots\ldots(4.1, 4)$$

and (4.1, 1) and (4.1, 2) are replaced by

$$\left.\begin{aligned}\sigma_z &= P\,.\,A/A_{\max},\\ u_z &= \frac{2(1-\sigma^2)a}{2}\cdot\frac{P}{E}\cdot\left\{C-\frac{D}{4(1-\sigma)}\right\}/A_{\max}\end{aligned}\right\} \quad \ldots\ldots(4.1, 5)$$

for either crack. It will be noted that if we do not impose this condition on $T_0\sqrt{(2c/a)}$ we cannot regard our solution, based on the pure elastic theory, as being in any agreement with the true solution at the ends of a crack based on interatomic or intermolecular considerations. For if T_0 or c is increased above the values for (4.1, 3) or (4.1, 4) to hold, then the maximum of our stress-strain curve is greater than the maximum theoretical stress P and so we cannot have a state of equilibrium and rupture must have commenced. Similarly, if T_0 or c is reduced, we cannot have equilibrium and the crack must be closing to some other shape. In fact, such a crack would close up completely unless some inclusion existed to maintain it.

Also it will be observed that conditions (4.1, 3) and (4.1, 4) are the conditions for the corresponding cracks to produce rupture. The conditions of (4.1, 3) and (4.1, 4) are, in fact, the equivalents of the usual criteria given by Griffith (1924).

Inserting the numerical values in (4.1, 3) and (4.1, 4), we have:—

For rupture,

$$\left.\begin{aligned}T_0\sqrt{c} &= 1{\cdot}269\,P\sqrt{a} \quad \text{for penny-shaped crack,}\\ T_0\sqrt{c} &= 0{\cdot}808\,P\sqrt{a} \quad \text{for a “ Griffith ” crack.}\end{aligned}\right\}. \quad \ldots\ldots(4.1, 6)$$

The form of the law of force derived is shown in figure 5; it is plotted in the most suitable non-dimensional form, σ_z/P against $\frac{E}{P}\cdot\frac{2u_z}{a}$, σ being put equal to 0·25.

We have thus far identified our law of stress-strain with a law expected from interatomic attractive forces only by the equivalence of the maxima, the slope at zero strain, and a diminution of the stress to zero for infinite strain. In fact we have determined the stress-strain law (assuming P is known) by this equivalence, as we had only the one parameter $T_0\sqrt{2c/a}$ in our expression for stress and strain. This makes us unable to obtain exact agreement between the stress-strain law derived as above and that expected from interatomic forces.

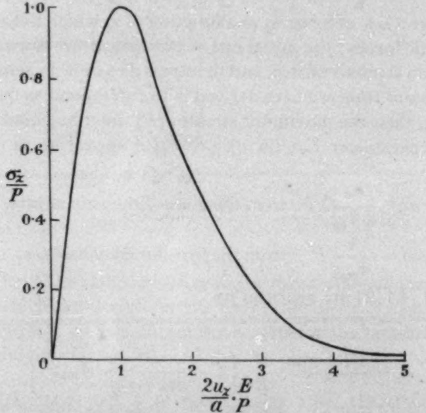


Figure 5. σ_z as a function of u_z.

The graph is plotted in the dimensionless form $\frac{\sigma_z}{P}$ against $\frac{2u_z}{a}\cdot\frac{E}{P}$.

There are two main differences:—Firstly, in the stress-strain law derived above, the slope increases with strain from E to $1\cdot 5E$ (approx.) before decreasing to zero at the stress maximum; this in turn causes the stress maximum to be attained for a smaller strain than that expected from interatomic laws of force. Secondly, the stress should tend to zero for large strain as $K/(2u_z+a)^3$, i.e., a van der Waals law of force (De Boer, 1936), while the stress-strain relationship found above gives $\tan^{-1}(\lambda u_z)$ which is a logarithmic decay. However, we know from the ideal elastic solution that σ_z would be zero on a surface only $\frac{1}{2}a$ from the first plane of atom centres, so that for large u_z (near centre of the crack) our model will not be greatly in error.

4.2. **One other result can be derived from our law of force.**

Our assumed law of force allows us to find the surface energy of the body. It is simply

$$\left.\begin{aligned} S &= \int_0^\infty \sigma_z \, d(u_z), \\ &= \int_0^{a/2} \sigma_z \, d(u_z) + \int_{a/2}^\infty \sigma_z \, d(u_z), \end{aligned}\right\} \qquad \ldots\ldots(4.2,\,1)$$

$$= I_1 + I_2. \qquad \ldots\ldots(4.2,\,2)$$

I_1 can be calculated numerically in each case using table 1, or from figure 5, once the values of P, E, σ and a are known. I_2 can be calculated, as a small correction, by using the fact that in this region σ_z is due to van der Waals's forces which die away as $K/(a+2u_z)^3$, K being determined by the value of σ_z at $u_z = \frac{1}{2}a$; I_2 is a correction of the order of 1% or less:

$$I_2 = a[\sigma_z]_{u_z=\frac{1}{2}a}.$$

Alternatively the integral may be found to sufficient accuracy by transforming it to the integral

$$S = \frac{(1-\sigma^2)P^2 a}{E(A_{\max})^2} \int_0^a A\, d\left\{C - \frac{D}{4(1-\sigma)}\right\}, \qquad \ldots\ldots(4.2,\,3)$$

which from figure 5 gives ($\sigma = 0{\cdot}25$)

$$S = 0{\cdot}92 \frac{P^2 a}{E}. \qquad \ldots\ldots(4.2,\,4)$$

This enables us to put our criterion for rupture in a form analogous to that of Griffith (1921),

$$T_0 \sqrt{c} = 0{\cdot}84 \sqrt{(S \,.\, E)}. \qquad \ldots\ldots(4.2,\,5)$$

4.3. Unfortunately the only results which allow full comparison of theory with experiment are those of Griffith, which are for glass—a non-crystalline substance. If, however, we assume that the crack is such that at least at its ends it is equivalent to a cut in a silica crystal along a basal plane and the mechanism of fracture is to cause cleavage to spread along this plane, we can use the above analysis. In fact, we assume that fracture begins with the spread of the edge of the crack through a crystallite of silica.

We then have

$E = 6{\cdot}34 \times 10^{11}$ dynes/sq. cm. from Griffith,

$S_{\text{expt}} = 546$ dynes/cm. from Griffith (by extrapolation from the liquid phase),

and the normal rupture strength for his glass

$$= 1{\cdot}83 \times 10^9 \text{ dynes/sq. cm.}$$

For quartz we have $a = 5 \times 10^{-8}$ cm. $= 1{\cdot}97 \times 10^{-8}$ in.

Griffith (1921) performed a series of experiments to determine $T_0\sqrt{c}$ for cracks produced artificially with a diamond. His values varied with annealing temperature, etc. (see Griffith, 1924), but by careful considerations of the factors involved, he finally found consistent values for $T_0\sqrt{c}$ of $1{\cdot}43 \times 10^7$, compared with $2{\cdot}69 \times 10^7$ of his earlier paper (T_0 in dynes/cm^2, c in cm.).

This gives us a value of $P = 8 \times 10^{10}$ dynes/sq. cm., a result which is slightly lower than that of $1{\cdot}3 \times 10^{11}$ given by Griffith (1924) by taking the radius of curvature of the end of the crack as 5×10^{-8} cm.

Using this value of $P = 8 \times 10^{10}$ dynes/sq. cm. we find that for the normal rupture strength the glass must contain cracks of half-length $c = 6{\cdot}1 \times 10^{-5}$ cm.

We may also use this value of P to find S, the surface tension. We find

$$S = 466 \text{ dynes/cm.},$$

which is somewhat lower than the value given by Griffith (1924).

§ 5. POSSIBLE PHYSICAL FORMS OF A CRACK

5.1. We may now examine our model to see whether we can avoid the mathematical fiction that when tensions are removed the crack is merely a plane across which no cohesive forces can act. For the above analysis to hold, it is sufficient that the unstrained state of the body should be that of a perfect solid lattice, and that for tensions sufficiently large for rupture, the crack surfaces should be free from forces (i.e. hydrostatic pressures, etc.). Our analysis does not forbid an inclusion that does not fill the space within the crack at rupture, and does not itself adhere strongly to the surfaces of the crack, i.e. a stress less than the rupture stress will separate it from the surface of the crack. This would obviously permit gaseous inclusions. Examination of the shape of the crack at rupture shows that the included atoms could at least extend to within two or three atom diameters of the "edge", and yet allow the surface to be "free" at rupture. If the included plane of atoms is such that the inclusion is solid or liquid, and does not adhere strongly, the only surface forces at rupture on the crack will be slight vapour pressures. If the included atoms are gaseous, the pressures may be greater.

5.2. An alternative hypothesis is to assume that one plane of atoms is missing between the faces of the crack, the penny-shaped hole being filled by a gas or other inclusion. Such an inclusion is necessary to maintain the hole; if the inclusion is not present, rough analyses show that cracks greater than about 10–20 atoms long close spontaneously or under thermal agitations. The hypothesis of such a crack does not greatly increase the size of crack required for rupture. This may be seen as follows :—

The stresses on the first plane of atomic centres in this case will be less than those given above as the plane is now distant a from the plane of symmetry.

Let us assume that to a first approximation it is equal to the stress calculated by the former analysis for the planes $z = \pm a$. Then

$$\left.\begin{aligned}(\sigma_z)_{z=a} &= \frac{2T_0}{\pi} A\left(\frac{a}{c}\right)^{\frac{1}{2}}, \\ &\simeq \frac{1}{\sqrt{2}} \cdot \frac{2T_0}{\pi} A\zeta_0^{-\frac{1}{2}}.\end{aligned}\right\} \quad \ldots\ldots(5.1)$$

Also we have no longer to equate the maximum value of this to P, the maximum cohesive stress which is theoretically possible for a perfect solid, for the atomic plane at distance a is missing in the crack itself. If we estimate the effect of this, using a suitable law of interatomic force, we find that if P_1 is the new maximum attractive force, then

$$P_1 \simeq 0{\cdot}8P. \qquad \ldots\ldots(5.2)$$

Hence we have for rupture

$$\left.\begin{aligned} P &\simeq T_0 \frac{2A_{\max}}{\pi} \zeta_0^{-\frac{1}{2}} \cdot \frac{0{\cdot}70}{0{\cdot}80}, \\ \text{or}\quad P &= 0{\cdot}487\, T_0 \zeta_0^{-\frac{1}{2}}. \end{aligned}\right\} \qquad \ldots\ldots(5.3)$$

For the glass used by Griffith, a crack about 3600 atoms long is now required to account for observed rupture stress.

As before, a "non-adhering" solid or liquid inclusion is permissible and has no appreciable effect on the analysis given above for rupture. We may also have a gaseous inclusion if it is not strongly absorbed by the material.

It will be seen that the effect on the rupture strength will be the same even if the inclusion has a less regular shape than that suggested above, i.e. a single plane of atoms, provided that it is only a single plane near the edges of the crack. "Near" in this case may be of the order of only 10–100 atom diameters, provided the inclusion does not broaden too rapidly beyond this distance.

ACKNOWLEDGMENTS

The author wishes to express his thanks to Professor N. F. Mott, F.R.S. and Mr. F. R. N. Nabarro for several illuminating and helpful discussions during the preparation of this paper, and to Imperial Chemical Industries for financial support.

REFERENCES

de Boer, J. H., 1936. *Trans. Faraday Soc.*, **32**, 10.
Elliott, H. A. and Sneddon, I. N., 1947. (In the press.)
Griffith, A. A., 1921. *Phil. Trans. Roy. Soc.*, A, **222**, 180.
Griffith, A. A., 1924. *Proc. Int. Congr. Appl. Mech.* (Delft), 55.
Inglis, C. E., 1911. *Trans. Inst. Nav. Arch.*, **55**, 219.
Sneddon, I. N., 1947. *Proc. Roy. Soc.* (In the press.)
Westergaard, H. M., 1939. *J. Appl. Mech.*, **6**, 49.

MULTIPLE-BEAM LOCALIZED FRINGES: PART I.—INTENSITY DISTRIBUTION AND LOCALIZATION

By J. BROSSEL,
Paris
(Now at Manchester University)

Communicated by S. Tolansky; MS. received 4 *September* 1946

ABSTRACT. The interference taking place when a parallel beam of light is incident on a wedge with highly reflecting surfaces is studied theoretically, an approximation correct to the third order being used and non-normal incidence considered. In addition to the well known plane of fringe localization close to the wedge, there is an infinity of other planes where clearly resolved fringes are observed even at distances up to several metres from the interferometer. In most cases the spacing of the maximum is a sub-multiple of that given by the classical formula $N\lambda = 2t \cos \theta$, but if minor details are taken into account the fringe period is the same in all planes of localization. This phenomenon is repetitive at equal intervals along the optical axis. All the effects have been experimentally observed. The significance of two fringe-broadening terms in the analysis is explained with reference to practical cases and the importance of small gaps is stressed. The theory predicts that a similar repetitive law of non-classical fringe spacing should exist at large distances along the line of greatest slope of the wedge, i.e. at large gaps of a centimetre or more.

§1. INTRODUCTION

Two types of low-order multiple-beam interference fringes have recently been extensively applied to the study of surface topography (Tolansky, 1943 a):

(*a*) "Fizeau" (localized) fringes of equal thickness formed by monochromatic light between two surfaces of reflection coefficient 85% to 95%.

(*b*) Fringes of equal chromatic order formed in a similar interferometer with white light, being channelled spectra with Fabry-Perot type intensity distribution (Tolansky, 1945).

The methods have ready application, for the fringes are easily obtained with high intensity and extreme sharpness.

It is the purpose of this paper to discuss the formation of these fringes and draw attention to novel properties predicted by theory and confirmed by experiment.

§2. PATH DIFFERENCE AFTER MULTIPLE REFLECTION

Figure 1 shows the arrangement used for Fizeau fringes. A monochromatic point source S is situated at the focus of a lens C and the resulting parallel beam falls on the interferometer I, which consists of two highly reflecting slightly transparent surfaces close together, on which fringes of equal thickness are formed. An objective M projects the fringes on to the plane E. If white light is used and a spectrograph slit brought into plane E, fringes of equal chromatic order appear in the spectral plane.

The case of a simple wedge is easily analysed mathematically, as shown below. For the corresponding experimental observations, two silvered Fabry-Perot flats were used with thin mica spacers.

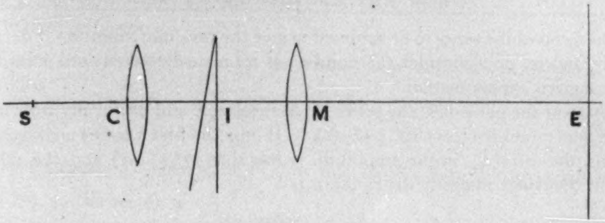


Figure 1.

Figure 2 represents a wedge, of which one reflecting plane is OY and the other is at a small angle ϵ to this; their line of intersection passes through O perpendicular to the plane of the figure. Let the intensity reflection coefficient of each plane be r. A plane wave approaches from the left at an angle of incidence θ and suffers multiple reflection inside the interferometer. This gives rise to a family of plane waves $\pi_0, \pi_1 \ldots \pi_n \ldots$, the combination of which defines the state of interference

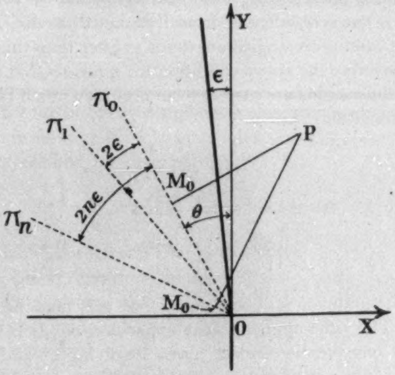


Figure 2.

at the point P (X_0, Y_0). These plane waves are in phase, have intensities decreasing geometrically, and the angle between successive members is 2ϵ.

The resultant vibration at P (X_0, Y_0) is given by

$$A \cos \Delta = \sum_n ar^n \cos \phi_n. \qquad \ldots\ldots(1)$$

This expression depends only on the position of P relative to the family $\pi_0, \pi_1 \ldots \pi_n \ldots$. The resultant at P is the same for all positions of the interferometer which do not affect the multiply-reflected wave-fronts $\pi_0, \pi_1 \ldots\, \pi_n \ldots$, i.e. states produced by the rotation of the wedge about its ridge. The path difference between the direct beam arriving at P and the beam which has suffered $2n$ reflections is

$$\delta_n = \mathrm{PM}_n - \mathrm{PM}_0 = X_0[\cos(\theta + 2n\epsilon) - \cos\theta] + Y_0[\sin(\theta + 2n\epsilon) - \sin\theta]. \qquad \ldots\ldots(2)$$

The terms of the series to be summed to give the resultant vibration A decrease rapidly, and we now consider the number of terms to be taken into account to obtain a given approximation.

Whatever the phase law, the series (1) is absolutely and uniformly convergent and its sum is always less than $a/(1-r)$. If only the first n terms are taken into account, the error R_n in the amplitude is less than $ar^n/(1-r)$ and the relative error in the fringe intensity distribution is

$$\frac{\Delta I}{I} = 2\left|\frac{R_n}{A}\right| \leqslant \frac{2ar^n}{(1-r)|A|}.$$

Hence if we wish to calculate the intensity with a relative accuracy better than one per cent, above the ordinate $I_0 = A_0{}^2$, we must choose n, so that

$$\frac{2ar^n}{(1-r)A_0} < \frac{1}{100}.$$

It is convenient to take for the value I_0 the minimum of the Airy distribution given by the particular phase law $\phi_n = n\psi$. It will be shown that the real phase law on the wedge is not very different from this, and that the intensity maxima obtained under the conditions studied are much greater than this particular value chosen for I_0. Therefore the above condition for n means that intensities in the neighbourhood of the maxima are correct to an accuracy much better than 1%.

To fulfil this condition,

$$r^n = \frac{1}{100} \cdot \frac{1-r}{2(1+r)},$$

we must take n greater than the following values

r	0·93	0·90	0·85
n	115	70	35

Returning to the summation of equation (1), we may take the finite number of terms as above. The same applies to the expansion of (2) in powers of ϵ; in addition we may conveniently neglect terms involving ϵ^4 and higher powers. Let the thickness of the wedge at the position $(0, Y_0)$ be t. Consider first the case of normal incidence $(\theta = 0)$. Then we have immediately

$$\delta_n = 2nt\left[1 - \frac{2n^2+1}{3}\epsilon^2\right] - 2n^2\epsilon^2 X_0. \qquad \ldots\ldots(3)$$

This formula is a good approximation. The absolute error in δ_n because of the neglected powers of ϵ is less than $\frac{2}{3}n^4\epsilon^4X_0$. For the representative case $\epsilon = 1$ minute of arc, $X_0 = 100$ cm., the error amounts to 5000 A. in the hundredth beam.

Equation (2) shows that the intensity distribution on the wedge ($X_0=0$) is not the same as that in Fabry-Perot fringes. For interference of low order (say less than 10) the difference is insignificant. Much of the work with this type of fringe has been carried out with such small gaps.

§ 3. FRINGES NEAR THE WEDGE

A. *Normal incidence*

From (3) we derive the law for the phase of the *n*th beam, i.e. where $\psi=\dfrac{4\pi t}{\lambda}$:

$$\phi_n = n\psi - n^2\frac{4\pi\epsilon^2 X_0}{\lambda} - n^3\frac{2\psi\epsilon^2}{3}. \qquad \ldots\ldots(4)$$

Near a thickness on the wedge of $N\dfrac{\lambda}{2}$ we may replace ψ by $(2N\pi+\alpha)$, α being less than 2π, and write (4) as

$$\phi_n = 2nN\pi + n\alpha - n^3\frac{2\epsilon^2}{3}(2N\pi+\alpha).$$

If we take two thicknesses differing by $\lambda/2$, the phase of the *n*th beam relative to the direct is in each case given by this formula. Similar expressions are only obtained if a term $\frac{4}{3}n^3\pi\epsilon^2$ can be neglected in comparison with 2π. Substitution of typical values shows that this is in general the case. The insignificance of this term shows that adjacent fringes always have nearly the same intensity distribution whatever their order.

Further, if the above term is negligible, then $\frac{2}{3}n^3\epsilon^2\alpha$ is also negligible compared with 2π. Hence the phase law becomes very nearly

$$\phi_n = n\alpha - n^3\frac{4\pi N\epsilon^2}{3}.$$

This shows that the intensity distribution changes slowly and periodically, and for the order of interference N it can be calculated as the square of the modulus of the following expression:

$$Ae^{i\Delta} = \sum_n ar^n e^{i\left(n\alpha - n^3\frac{4\pi N\epsilon^2}{3}\right)} = \sum_n ar^n e^{i(n\alpha-\beta\ )}.$$

It is seen from this that the total energy contained in a fringe cycle of 2π is not dependent on the phase relation β_n of the different component vibrations. Hence the main maxima of intensity of a fringe system are all greater than the miminum value of a system obeying the above law with $\beta_n=0$ (i.e. a Fabry-Perot type distribution) as assumed above.

If the phase terms β_n are small, the fringes are asymmetrical. As they become greater, secondary maxima occur in the interference pattern close to the main maxima. In all cases the main period is $\lambda/2$.

B. *Non-normal incidence*

Now consider the case of non-normal incidence. Expression (2) shows that a rotation θ of the axis results in another formula identical with the one for normal incidence. It is therefore obvious that the interference pattern in the plane π_0 for incidence θ is the same as that on the wedge surface for normal incidence. π_0

constitutes the Feussner *surface of localization* (Feussner, 1927). For large values of the thickness, π_0 can be at a great distance from the face of the wedge when the incidence is not critically normal.

It is easily shown experimentally that rotation of the interferometer about the ridge of the wedge, without changing any other part of the system, leaves the appearance of the fringes unaltered. At rotations greater than 20° the fringes split into two polarized components owing to the differential phase change at reflection with angle (Tolansky, 1944).

§4. ASYMMETRICAL BROADENING OF FIZEAU FRINGES

Expanding formula (2) in powers of ϵ for the case $\theta = 0$ we find that on the wedge, using similar approximations,

$$\delta_n = 2nt\cos\theta\left(1 - n\epsilon\tan\theta - \frac{2n^2+1}{3}\epsilon^2\right). \qquad \ldots\ldots(4)$$

This is the equivalent of the familiar formula quoted for the special case of maxima ($\delta_n = N\lambda$),

$$2t\cos\theta = N\lambda.$$

It is clear from expression (4) that the fringes on the wedge are sharper for normal than for non-normal incidence. The divergence of the phase law from that of Fabry-Perot fringes is, in the former case, $\frac{2}{3}n^2\epsilon^2$, and in the latter it is $n\epsilon\tan\theta$.

Consider the case of normal incidence with a pin-hole source of radius θ. When θ is of the same order of magnitude as ϵ, then any fringe-broadening due to the source extension arises from the $\cos\theta$ term in expression (4) and is significant only for large thicknesses. In effect the broadening is due to the different fringe spacing for each radial zone of the source about the axis. If $\Delta i/i$ is the fractional order broadening at the order of interference N which can be tolerated, the angular diameter of the source must be less than

$$\sqrt{\left(\frac{8}{N}\cdot\frac{\Delta i}{i}\right)}.$$

When the interferometer has a gap of some millimetres, the order of interference is so high that even very small source extensions show broadening, and this broadening appears as a square-topped intensity distribution (figure 3, II).

For gaps less than, say, 100 wave-lengths the same value of $\Delta i/i$ is obtained only for very large extensions (θ) of the source (up to 40°). The broadening is then partly due to the $\cos\theta$ term of equation (4) and partly due to the term $n\epsilon\tan\theta$. The fringes formed by a radial zone θ from the axis are not only displaced towards the side of greater gap but are also broader. Low-order fringes taken with poorly collimated incident light (i.e. non-parallel) exhibit a wing on the side of greater gap (figure 3, III).

This characteristic has found application in the interpretation of surface relief. Figure 4 shows fringes formed in the air film between an optical flat and a sheet of mica, both silvered, and pressed close together. In the first case (figure 4 *a*) strict collimation conditions are obeyed and the fringes have an intensity distribution similar to Fabry-Perot fringes (figure 3, I). When the pin-hole is replaced by an

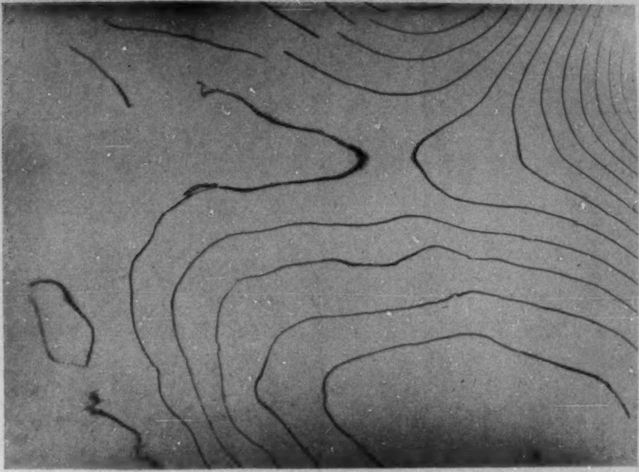


Figure 4 *a*.

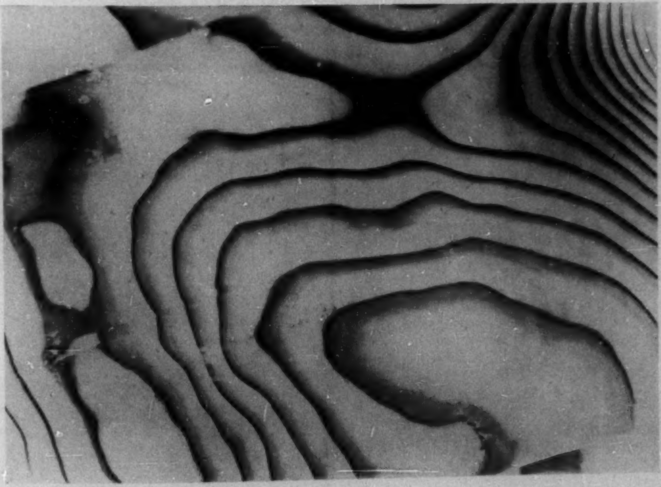
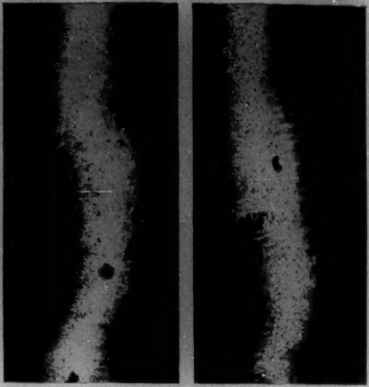


Figure 4 *b*.

Figure 5.



(a) (b)

Figure 6.

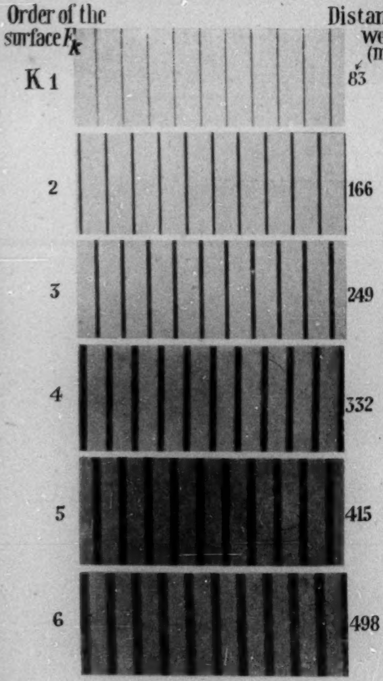

Order of the surface F_k
K 1
2
3
4
5
6
83
166
249
332
415
498

Figure 7.

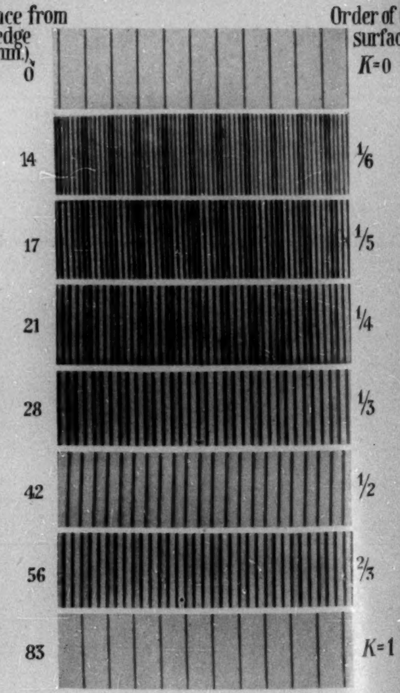

Distance from wedge (mm.)
0
14
17
21
28
42
56
83
Order of the surface K
K=0
1/6
1/5
1/4
1/3
1/2
2/3
K=1

Figure 9.

extended source the asymmetry of figure 3, III appears and the hills and valleys of the surface are clearly differentiated (figure 4*b*).

Any lack of parallelism in the beam results in an asymmetry of this type.

Wings are sometimes observed which seem to be attributable to diffraction and scatter at surface scratches; this is an unavoidable type of collimation error. Good flats usually have a perfect surface polish, and Fizeau fringes obtained between a pair, in a well collimated beam, are perfectly symmetrical and continuous. When one flat is badly scratched the fringes can never be obtained without a certain amount of asymmetry. A good example is shown in figure 5—a photograph kindly supplied by Mr. W. L. Wilcock of these laboratories. The fringes are formed between a polished diamond surface and an optical flat. This photograph was taken under most careful collimation conditions, but scratches on the diamond have caused an asymmetry of the fringes, the side of the fringe towards increasing gap being diffuse. A peculiar phenomenon observed with scratched surfaces is that the fringes with large wedge angles may occur as discontinuous series of dots. The effects are the same with fringes of equal chromatic order, which exhibit the wing on the blue side of the fringe.

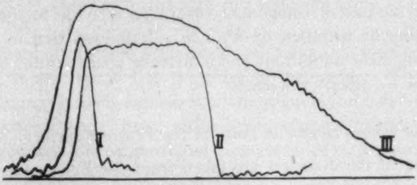


Figure 3.

Figure 6 shows a similar phenomenon with a very small wedge angle. The width of the fringe in the photograph is two per cent of the distance between orders.

In (*a*) the viewing microscope is focused on the plane of the silver. Surface scratches and "pin-holes" in the silver give bright points in the photograph. The same fringe is shown in (*b*) with the microscope at a slightly different focus position. Many of the bright points are now replaced by dark patches surrounded by white diffraction haloes, clearly seen on the side of greater gap (right-hand side). The changes of appearance depend very critically on the position of the microscope, and the size of the details so revealed bears no relation to the real size.

It is of interest to point out that, of the crystal cleavage surfaces so far studied in this laboratory, mica has always given perfectly smooth symmetrical fringes while diamond, calcite, selenite and baryta surfaces have exhibited the characteristics mentioned above. Possibly the effect is due to the existence of an unresolved cleavage structure.

A further cause of fringe asymmetry, arising from light unavoidably off collimation, is the scatter and unwanted reflections from the surfaces of all the optical components. This leads only to very weak wings, of intensity comparable with the normal background between maxima. This has had practical importance only when studying secondary fringe structure.

§ 5. INTENSITY DISTRIBUTION AWAY FROM THE WEDGE

Formula (3) explains an experimentally observed peculiarity of Fizeau fringes. If the collimator has a small pin-hole or a small slit and the interferometer is a simple wedge with its ridge parallel to this slit, then sharp fringes can be observed in many planes other than the Feussner surface.

A. *The interference pattern along the X-axis*

In normal incidence the phase, as shown above, is sufficiently nearly

$$\phi_n = n\psi - n^2 \frac{4\pi\epsilon^2 X_0}{\lambda} - n^3 \frac{2\psi\epsilon^2}{3}. \qquad \text{......(4)}$$

This formula also applies to non-normal incidence if X_0 is replaced by x_0, which is the rotated axis perpendicular to the incident wave front.

Taking two planes parallel to π_0 and a distance apart of

$$x = \mu \frac{\lambda}{2\epsilon^2} \quad (\mu = 0, 1, 2, \ldots.) \qquad \text{......(5)}$$

we see that the families of component vibrations arriving at these planes differ in phase, member by member, by $n^2\mu \,.\, 2\pi$. It follows that in any two planes parallel to the incident wave-front and a distance apart defined by (5), the interference patterns are exactly the same.

B. *Recurrence of fringes similar to those on the Feussner surface*

In particular, if the distances are taken from the Feussner surface then the observed pattern in each plane is identical with that on π_0 itself.

In a series of experiments designed to observe as many as possible of the orders (μ) of the surfaces of localization, two silvered flats were used. A large value of ϵ was adopted, there being 205 fringes in a field of diameter 43·2 millimetres with the mercury line 5461 A.

Formula (5) gives, from these figures, the value $x = 16{\cdot}5$ cm.

When a microscope was focused first on the fringes localized on the flats themselves and then successively on the planes x, $2x$, $3x$, ... $10x$, closely similar fringe systems were observed, the appearance of the field being quite different away from these positions. The only exception was that exactly in the middle of the above locations a similar pattern again occurred (i.e. at 8·2 cm., 16·5 cm., etc.). This property is general because on planes distant $\frac{x}{2} = \frac{\lambda}{4\epsilon^2}$ apart the interference patterns are the same, apart from a half-order displacement.

The proof is as follows:

Let ϕ_{1n} be the phase relation on the first plane and ϕ_{2n} be that on the second. We have

$$\phi_{2n} = \phi_{1n} - n^2\pi.$$

Now change the origin of ψ in equation (4) by π for the first surface only. With this new origin we find the phase relation on this surface becomes $\phi'_{1n} = \phi_{1n} - n\pi$ if a term $2\pi n^3\epsilon^2/3$ is neglected (as was done before in showing that adjacent fringes in a pattern have practically the same intensity distribution). Obviously ϕ_{2n} and ϕ' are identical because n and n^2 are even or odd together.

This half-order fringe shift is difficult to demonstrate experimentally. In the present case it was 0·1 mm. for positions of the microscope 10 cm. apart. In practice, therefore, it is convenient to consider that the interference repeats at intervals of $K\frac{\lambda}{4\epsilon^2}$, and these planes of repetition may conveniently be designated as the surface of localization of order K, say F_K. Figure 7 shows photographs of the patterns observed in the planes where K is integral, the exact values being indicated in the figure. Displacements of half an order have been given to the even-number patterns to accord with the above theory.

Patterns with spacing different from this were observed between these positions and will be described below. The effect of the extension of the source will be considered first.

If a perfect point source in the focal plane of the collimator is moved slightly off the axis in a direction parallel to the ridge of the wedge, the fringes in F_K are merely displaced parallel to their length and the appearance remains unaltered.

If the point source moves so that the incident plane wave front π_0 rotates through $d\theta$, remaining parallel to the ridge of the wedge, the fringes in F_K rotate about the apex of the wedge through the angle $d\theta$, their linear displacement then being $x_K \, . \, d\theta$.

The photographs in figure 7 were taken with a small slit source in the collimator arranged with its axis parallel to the ridge of the wedge. The square-topped intensity distribution is easily explained on the grounds of the above considerations. On a given surface F_K the fringe width is proportional to the source width $d\theta$, and for a given source it is proportional to the order K of the surface. The latter effect is apparent in figure 7.

C. *Differently spaced fringes in other planes of localization*

Between the surfaces with integral values of K, other positions are found where well-defined fringe systems occur. For example, we may take the half-way positions. The intensity is given by the square of the modulus of

$$A_{\frac{1}{2}}e^{i\Delta_2} = \sum_n ar^n e^{i\left(n\psi - n^2\frac{\pi}{2}\right)}.$$

Taking the real part of this expression, the summation becomes

$$A_{\frac{1}{2}}\cos\Delta_2 = \sum_p ar^{2p}\cos p(2\psi) + \sum_p ar^{2p+1}\sin(2p+1)\psi.$$

Each of these series can be calculated from the trigonometrical formula

$$\cos\eta + \rho\cos(\eta+\zeta) + \ldots + \rho^n\cos(\eta+n\zeta) + \ldots = \frac{\cos\eta - \rho\cos(\eta-\zeta)}{1-2\rho\cos\zeta+\rho^2}.$$

Finally, for the real part,

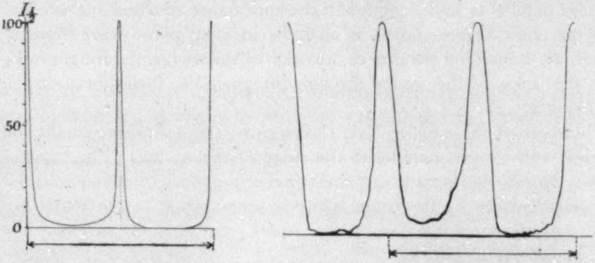
$$A_{\frac{1}{2}}\cos\Delta_2 = a\,\frac{1-r^2\cos 2\psi + r(1+r^2)\sin\psi}{1-2r^2\cos 2\psi + r^4}.$$

A similar calculation applies to the imaginary part, and the following table gives the resulting numerical intensity values for $r = 0{\cdot}92$. Figure 8 (*a*) gives the curve, and figure 8 (*b*) is a direct photometer record taken across a photograph of the fringes.

ψ	0°	5°	10°	20°	90°	180° $-\xi$	180°	180° $+\xi$	190°	200°	210°		270°		360° $-\xi$	360°
$I_{\frac{1}{2}}$	100	53	21	7·2	0·32	53	100	46·5	14	3·6	1·3	0	$2\cdot10^{-3}$	0	46·5	100

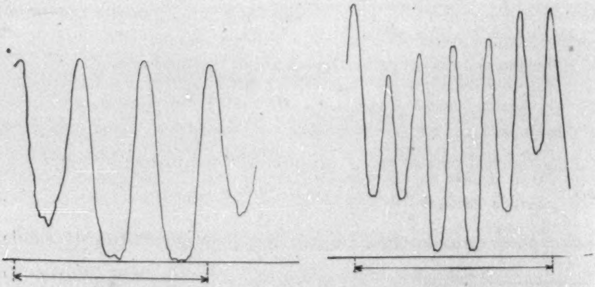
$$\xi=\frac{1-r^2}{2r} \quad \text{and for } r=0{\cdot}92,\ \xi \approx 5^\circ.$$

Sharp maxima occur with a spacing half that of the fringes on the Feussner surface. The minima, however, alternate in intensity so that the true period of these fringes is the same as the period of those on the wedge. Attention is drawn to the very low intensity of one background compared with the Airy distribution.



(a) $K=\frac{1}{2}$ calculated.

(b) $K=\frac{1}{2}$ record.



(c) $K=\frac{1}{3}$ or $\frac{2}{3}$.

(d) $K=\frac{1}{6}$.

Figure 8.

A similar calculation gives the intensity distribution at intervals $\frac{1}{3}$ and $\frac{2}{3}$ of the interval between the surfaces F_K (K integral). The fringes obtained have one-third of the spacing between maxima, and every third background is relatively intense (figure 8(c)). Figure 9 shows other intermediate cases at K values of $\frac{1}{4}$, $\frac{1}{5}$, $\frac{1}{6}$. All these photographs were over-exposed to record the background intensities. The microphotometer record for $K=\frac{1}{8}$ (figure 8(*d*)) shows that the maxima do not have equal intensity, two adjacent maxima of each cycle being quite intense, and exhibiting a high intermediate background. As the wedge is approached from this position (i.e. $K \rightarrow 0$) this characteristic becomes more dominant, until the Airy distribution is reached.

The intensity distribution can be calculated in the more general case of $K=1/m$ (m being any integer).

Assuming the phase law $\phi_n = n\psi - n^2\frac{\pi}{m}$, the intensity is given by the square of the modulus of

$$A_{1/m}e^{i\Delta_m} = 1 + \sum_{q=1}^{q=m} \left\{ \left(re^{i\Psi}e^{-iq\frac{\pi}{m}} \right)^q \frac{1}{1-(-1)^m r^m e^{im\Psi}} \right. .$$

This phenomenon of multiple localization occurs with many types of multiple-beam interferometer, and recalls the properties of the zone plate. In fact the experimental arrangement of the wedge described above is similar to Fresnel's mirror or the Billet split lens, being merely the multiple-beam version of it. With the two beam interferometers the fringes always have a $\sin^2$ intensity distribution and the order spacing is always given by the simple formula $\delta = N\lambda$. For the case of multiple beams, the fringes appear as very sharp lines in most cases, but the distance between the lines is given by the above formula in only a few instances. If, however, it is borne in mind that the true fringe cycle is determined by secondary phenomena (fringe background and inequality of maxima) it is easily seen that the true period of the fringes always obeys the relation $\delta = N\lambda$.

D. *Recurrence of the interference pattern along the wedge*

We have just examined some of the consequences of the phase law

$$\phi_n = n\psi - n^2\frac{4\pi\epsilon^2 X_0}{\lambda},$$

in which

$$\frac{4\pi\epsilon^2 X_0}{\lambda} = K\,.\,2\pi;\ \pi;\ \frac{\pi}{2}.$$

On the wedge itself the phase law is more exactly given by

$$\phi_n = n\psi - n^3\frac{8\pi t\epsilon^2}{3\lambda},$$

as seen from equation (4). This term in n^3 gives the same effects as noted above for the n^2 term when it is equal to $2N\pi$, π, $\pi/2$, because n^2 and n^3 are even and odd together and the amplitude series are identical for the two cases.

Hence at the thicknesses of the wedge given by

$$t = N\,.\,\frac{3\lambda}{8\epsilon^2}$$

a Fabry-Perot type intensity distribution should exist.

An interesting deduction is that between two thicknesses differing by $3\lambda/4\epsilon^2$ the number of fringes observed should be one less than that predicted by the simple theory.

Tolansky (1943) has noted a similar deviation of fringe spacing from the predictions of simple theory in the case of non-localized Fabry-Perot rings (Tolansky, 1943 b).

Another important point is that at thicknesses

$$t=\frac{3\lambda}{16\epsilon^2}+N\frac{3\lambda}{8\epsilon^2}$$

fringes should be observed which are bright lines on a dark background and have a distance between them only half that given by the simple theory.

It is proposed to investigate this prediction experimentally in this laboratory in the near future.

ACKNOWLEDGMENTS

I wish to express my thanks to Dr. S. Tolansky for his interest and encouragement. Among my co-workers, I am particularly indebted to W. K. Donaldson, who first constructed much of the apparatus necessary for this research.

REFERENCES

FEUSSNER, 1927. Gehrcke's *Handbook der Physik—Optik*, vol. i.
TOLANSKY, 1943 a. *Nature, Lond.*, **152,** 722; 1943 b. *Phil. Mag.*, **34,** 55; 1944. *Ibid.*, **35,** 120; 1945. *Ibid.*, **36,** 225.

MULTIPLE-BEAM LOCALIZED FRINGES: PART II.—CONDITIONS OF OBSERVATION AND FORMATION OF GHOSTS

BY J. BROSSEL,
Paris
(Now at Manchester University)

Communicated by S. Tolansky; *MS. received* 4 *September* 1946

ABSTRACT. Factors which tend to reduce the perfection of the fringes exist in the optical systems associated with a Fizeau multiple-beam interferometer and in its complexity of surface detail. A parallel is drawn between the mechanism of image formation of these fringes and the theory of microscope image-formation due to Abbe. Experiments are described showing the effect of the presence of diaphragms in the second focal plane of the objective and also in contact with the objective. The conclusion reached experimentally is that if the objective has insufficient aperture the fringes broaden and secondary maxima appear.

At the same time, the resultant of a non-infinite series of geometrically decreasing waves is studied. The results obtained are also valid for the Lummer-plate interferometer.

Detailed attention is given to the parasitic reflected light from the many glass-air surfaces of the optical train used, and the resulting ghosts are described.

It is shown that the interferometer gap must be of only a few wave-lengths if the localized fringes are to follow the contour lines correctly over the smallest surface details.

§ 1. INTRODUCTION

In this paper an examintion is made of the necessary conditions which must be fulfilled in the observation of multiple-beam Fizeau fringes in order to obtain a final image identical with the primary interference pattern. The investigation was undertaken when the question arose as to the limit of surface detail which can be resolved by the technique, and the accuracy with which the fringes follow surface contours.

It is assumed throughout that the order of interference is very small, that the idealized case of a simple wedge is under consideration, and that the fringes are those localized in the Feussner surface of zero order.

§ 2. EFFECTS OF THE PRESENCE OF DIAPHRAGMS ON THE STRUCTURE OF THE IMAGE

(a) *Abbe type experiment*

The present observations resemble the phenomena studied by Abbe in developing his theory of the resolving power of the microscope and, in particular, his case of a grating (R) in a parallel beam (figure 1 (*b*)) and the conditions for similarity of the image to the object. Comparing the optical arrangement used here (figure 1 (*a*)) with that of Abbe, it is apparent that the only difference lies in the method of division of the wave. Abbe has clearly shown that the image E of the grating R is an interference pattern arising from the reciprocal grating R_1 in the focal plane of the objective. Artificial modifications of R_1 produce false detail in the image.

In the interferometer case there is division of amplitude of the incident beam, and the emergent family of multiple reflected wave fronts $\pi_0, \pi_1, \ldots, \pi_n$ (see Part I; Brossel, 1947) give a set of equally spaced images of the original point source in the focal plane of the microscope objective M. The intensities of these images decrease geometrically, and there is a definite phase relation between them. For convenience these will be referred to as the grating I_1, reciprocal to the interference phenomenon I. The fringes formed on E can be calculated from this grating I_1: any defect of phase or amplitude imposed on the grating I_1 will cause fundamental modification of the final image of the fringes.

The functioning of many multiple-beam interferometers can be explained in terms of this grating—in particular the interferometers of Fabry-Perot and Lummer-Gehrcke. In Part I it was indicated that Fizeau fringes under discussion are the multiple-beam version of the two-beam fringes of Fresnel's mirrors or Billet's split lens. The relationship is clearly in evidence from figure 1. These fringes, E, could be considered, too, as the different orders of spectra coming from an echelon. The plate thickness would be equal to the thickness of the interferometer on the optical axis, but the angular dispersion would not be due to diffraction (there is no amplitude factor dependent on direction).

These multiply-reflected images of a point source were used by Fabry (1927) in the determination of reflection coefficients. They served the same purpose in this investigation, but a photographic method was adopted in place of his polarization technique. Their use more generally in photographic photometry has been overlooked in the past, but they form the only true logarithmic intensity scale. The source does not require stabilizing, but the reflectivity usually has to be evaluated by some auxiliary method.

The phase relationship between the elements of the grating I_1 can be calculated by a simple geometrical method. Translation of the interferometer leaves the positions of the images I_1 unchanged, but modifies their phase relationship. When the wedge is in the focal plane of the objective, with its apex on the axis, all the sources I_1 are in phase and the resultant final fringes at infinity are the maxima of a grating. A translation in the plane of the figure perpendicular to the optical axis introduces a constant phase difference between adjacent sources which ensures the corresponding translation of the final image pattern.

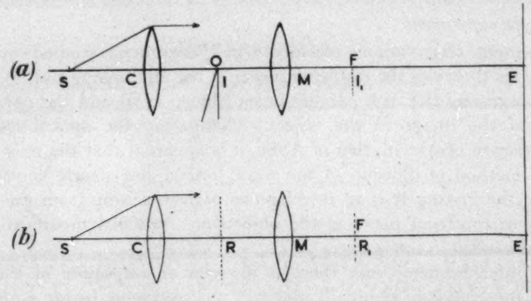
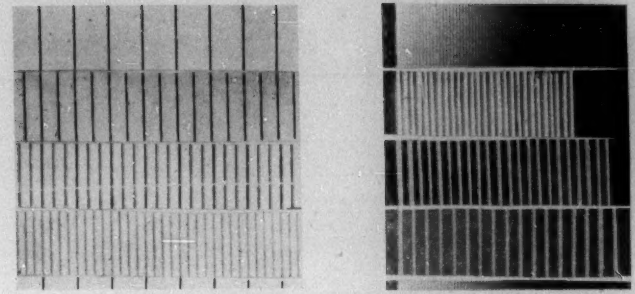


Figure 1.

When the interferometer I is at a distance $\overline{MO} = u$ from the objective, the path difference between the first and the *n*th source is $2n^2\epsilon^2(f-u)$. (The notation of Part I is used throughout.) It is possible to choose the experimental conditions so that the supplementary paths involved exactly compensate this value. It is sufficient for this purpose to place the screen E so that $\overline{ME} = v = \frac{uf}{u-f}$, i.e. in the plane conjugate to I. The image of the fringes can be very close to the grating I_1, from which they arise, and yet remain perfect.

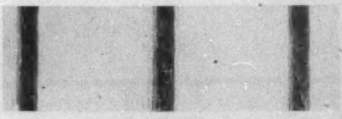
In planes other than this conjugate, the multiple localization phenomena of Part I are observed, but it is apparent that they may also be considered as arising from the grating I_1.

Consider now the effect on the fringes at E, arising when various obscuring diaphragms are placed in the plane I_1. For instance, alternate images can be masked and the resulting grating is then equivalent to that given by a wedge of twice the angle, and having a reflectivity r^2 instead of r. The result is that the fringes observed with such a mask will have a spacing only half that of the

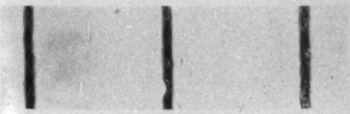


(a) (b)

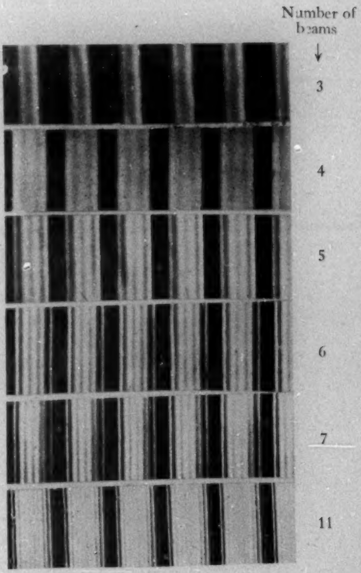
Figure 2.



(a) Number of beams 30.



(b) Number of beams ∞.


Number of beams
↓
3
4
5
6
7
11

(c)

Figure 3.

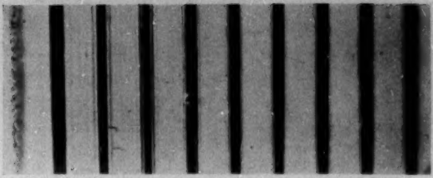


Figure 5.

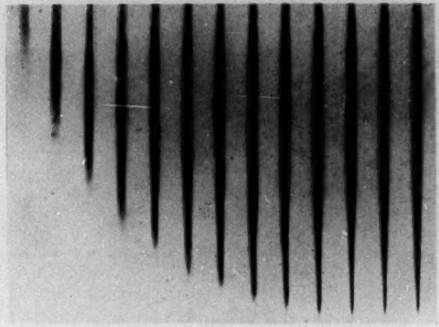


Figure 6.

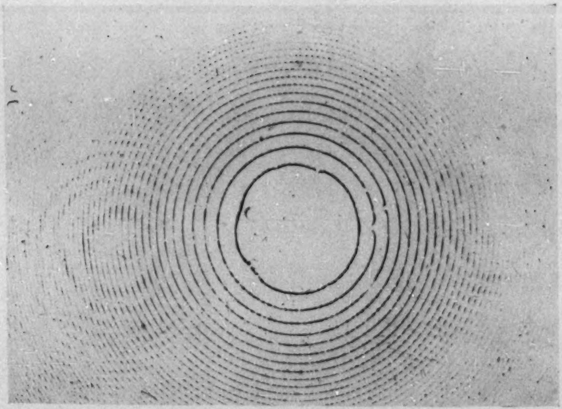


Figure 8.

original fringes, and must be broader. This corresponds exactly with observation, and figure 2 shows the result of a series of experiments of this type.

The masks, shown in figure 2 (*b*) (excluding top and bottom pictures), were suitably adjusted in the focal plane I_1 of the objective, and the corresponding fringes observed are shown opposite each in figure 2 (*a*). To complete the picture, the top photograph on the left shows the unmodified fringes, and at the same level at (*b*) are shown the unmodified multiple images in plane I_1. (The source was not a pin-hole but a small slit.) When two out of three, three out of four, etc., images are masked in I_1 the observed fringes are three times, four times, etc., as close together.

The multiple images were first recorded on photographic paper which was then suitably cut with a razor blade to make the mask. This was then adjusted in the focal plane I_1 with the aid of a micrometer screw and a high magnification simple eyepiece. The projection objective used to form the fringe image was a Zeiss anastigmat, f/4·5 and 50 cm. focal length.

(b) *Secondary maxima*

A second set of experiments was performed in which the series of images was restricted to a given number, the less intense tail of the diminishing series being masked. This is a convenient method of studying the combination of a non-infinite series of waves of diminishing amplitude. This case is met with in practice when the objective has an insufficient aperture, as detailed below. The results also apply to the Lummer plate, which seldom uses more than 30 beams. If the grating I_1 is restricted so that only the first two sources are uncovered by the mask over the series tail, the fringe images have a $\sin^2$ intensity distribution. For three sources (the three-line grating) the principal maxima are sharper than in the previous case and there is one secondary maximum between successive orders. With four sources there are two secondary maxima, and so on. The development of the phenomenon was followed in detail with an increasing number of sources. The number of secondary maxima is identical with that in the classical grating case of Abbe, but the intensities are influenced by the fact that in the present case a logarithmic decrease of intensity obtains. With a reflection coefficient of 0·92 and 20 sources the 18 secondary maxima were easily photographed. Beyond this the secondary maxima half-way between principal maxima were lost in the background, but the secondaries close to the principal maxima were resolved when as many as 40 sources were used. Figure 3 (*a*) shows a photograph taken with 30 sources, and four or five secondaries are apparent. Figure 3 (*b*) shows the same fringes with an "infinite" series of sources (Fabry-Perot distribution). A number of intermediate cases are given in figure 3 (*c*), the number of exposed sources being printed to the right of each photograph. (The fringes are black.)

The intensity distribution in the fringes given by 30 sources has been studied by photographic photometry and the curve obtained is given in figure 4. These measurements apply to an order of interference of 25. The term $\frac{4}{3}n^3\epsilon^2 t$ is important enough to cause a distinct asymmetry (see Part I). On the scale given, the peak of the maximum (omitted) has intensity 100.

As a result of this it appears that a multiple-beam interferometer should utilize at least 40 beams. For instance, the length of a Lummer plate should be at least 80 times its thickness. If this condition is not fulfilled, great care should be taken when the intensity and position of a hyperfine structure component close to a strong maximum and less than 10 % of its intensity are being measured. These data apply to a perfect instrument used with very sharp lines, and the tolerances are often greater in practical cases.

(c) *Effect of the aperture of the objective.*

Returning to the conditions desirable for observing Fizeau fringes, it is immediately apparent from the above that the aperture of the projecting lens which collects the beams is a fundamental consideration. When this is insufficient it acts as a mask and the fringe image is broadened, whilst secondary maxima appear. The beams leaving the interferometer all fall on the same side of the optical axis. If the lens diameter is less than the illuminated field of the interferometer, all the beams (or, say, 100) on one edge of the interferometer can pass through the objective: the corresponding fringe image will have the Fabry-Perot distribution. At the opposite edge of the interferometer field only one or two of the beams leaving the interferometer pass through the objective: the corresponding fringe image will then have a sine-squared distribution. Between these two extremes all the intensity distributions described above are met with; in traversing the field from the broad fringes, the principal maxima become sharper and sharper, the secondary maxima increasing in number. Finally, these secondaries are lost in the background and the region of sharp fringes is reached.

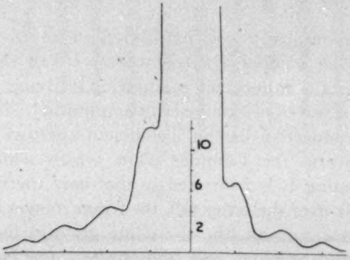


Figure 4.

A photograph of this effect is given in figure 5. The diaphragm across the objective was a slit 4 cm. wide, with its edges parallel to the fringes. The field of the interferometer was 6 cm. in diameter.

Similar effects are also observed with a circular diaphragm, but they differ in that the intensity distribution may vary along the length of a fringe. The photograph in figure 6 was taken under similar conditions to that of figure 5, the rectangular objective stop being replaced by a circular hole of 4 cm. diameter.

In the microscope normally used to observe Fizeau fringes the field diaphragm is small enough to prevent such an effect. Occasionally the fringe definition at one edge of the field is observed to be imperfect.

An alternative to modifying the amplitudes of the sources I_1 is the modification of their relative phases.

A random variation in phase is readily imposed by introducing a plate of glass in the plane I_1. The definition of the observed fringes is then materially affected.

If a vessel constructed like a hollow echelon were placed correctly in the plane I_1, the instrument would behave as a refractometer—being in fact the multiple-beam version of the Rayleigh instrument. Of course it is much simpler to introduce the dispersive medium between the plates of the interferometer itself and use gaps of the order, for instance, of centimetres.

§ 3. GHOST IMAGES IN FIZEAU FRINGES

(a) *Glass surfaces before the interferometer*

In studying the secondary maxima it was found that weak systems other than those mentioned above are present. These "ghosts" have a comparable intensity to the secondary maxima of interference—in the region of 4% of the main maxima. They are due to parasitic reflections from the unsilvered glass surfaces. 90% of the light incident on the interferometer is reflected by the first silvered surface and 4% of this is returned towards the observer from each glass surface encountered. Each of these surfaces therefore produces a ghost point-source which has its own set of fringes.

The optical arrangement used is shown in figure 7; each surface is numbered, and a monochromatic filter is usually included (3, 4). The plane surfaces (3 and 4) of the filter and the unsilvered surface (5) of the first optical flat give ghost point-sources s_3, s_4, s_5 at infinity and in general these do not lie on the axis. As a result (because of the $\cos\theta$ law) the corresponding ghost fringes have each a distance between orders slightly greater than the main system, and appear as a fringe structure on the side of greater gap. The resolution of this structure improves with increasing order. These ghosts can be eliminated by increasing the tilt of the offending surfaces, or by fitting a screen in the second focal plane of the objective so that only the main set of multiple images passes.

Surfaces 1 and 2 of the collimating lens give two ghost point-sources s_1, s_2 at finite distances. The ghost fringes are those formed in a wedge with a point-source at a finite distance—a much more complicated situation than the parallel beam analysed above. It can be shown that locally the fringes formed are the same as if the source were moved to infinity along the actual direction of incidence. At the foot of the perpendicular to the wedge dropped from s_1 (or s_2) the ghost and the main fringe coincide, and away from this region in any direction (along the fringe too) the ghost moves to the side of greater gap of the main fringe. This has been observed experimentally.

It is difficult to eliminate the ghosts 1 and 2. The collimating lens may be tilted a little and the interferometer placed at a distance such that the ghost beams are not collected. It seems that the best solution is to use a plano-convex lens with its plane side towards the interferometer. Ghost 2 is then in the same class as 3, 4 and 5 above and the real point-source s_1, if troublesome, can be brought to a focus on the interferometer in an unimportant area.

(b) *Surfaces following the interferometer*

The ghosts arising from surfaces 6, 7 and 8 on the exit side of the interferometer are due to the reflection in each case of 4% of the incident energy, and of this 90% is reflected again by the second silvered surface towards the observer. These ghosts are images of the fringes given by the mirrors 6, 7 and 8 in combination with the plane-silvered mirror and the objective. They are not localized on the interferometer and can therefore exhibit great complexity since the plane of observation is not the true plane of localization. In these experiments the ghost 6 is localized at about 3 cm. behind the main interference pattern. The surface of the microscope objective (7) is plane, and with the usual focal lengths its ghosts fringes are localized at 7 to 10 cm. behind the interferometer. With the small dispersion used (angle of the wedge a few minutes of arc) the pattern away from the surface of the wedge does not change rapidly with X (Part I—Intensity distribution away from the wedge) and ghost 6 always appears simply as an out-of-focus fringe superposed on the main system. Ghost 7 is seldom observed. Surface 8 produces an inverted image of the fringes, which always seem to be localized close to the interferometer. A bloomed objective appears to offer the only improvement. This ghost is easily identified when using Fizeau fringes; as the field of the interferometer is explored under the microscope, the ghost and the main fringe system move in opposite directions.

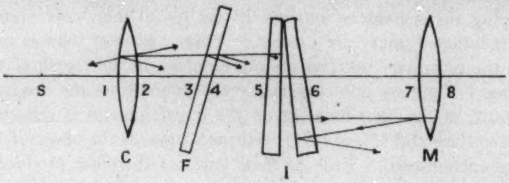


Figure 7.

(c) *False details arising from the ghosts*

These ghosts from the surfaces of the optical components frequently cause false detail to appear in the fringes. It must be remembered that when studying crystal topography with a microscope there can be more unsilvered glass surfaces than those described above, and each gives some form of ghost. The most troublesome type of ghost arises when two surfaces close together are nearly parallel. The result is that the intensity of the main fringes is modulated by the ghost pattern. This effect in multiple-beam Fizeau fringes is well known, having been formerly described, and is mentioned here for the sake of completeness. A photograph clearly demonstrating the effect is given in figure 8, and shows multiple-beam Newton's rings with two $\sin^2$ ghosts modulating the main fringe system. It is rarely that the bright parts of the main fringes can be joined up to give a simple secondary system as here.

§4. EFFECTS OF SURFACE STRUCTURE

In conclusion it is apparent that large-aperture objectives must be used to project Fizeau fringes if a sufficient number of the interfering beams are to be collected. Further, since the objective is collecting only plane or nearly plane waves, it is working under the worst conditions for the resolution of the true details of the surface, as shown by Abbe. This applies to the resolution in directions perpendicular to the axis. The Fizeau method, however, is used only to reveal details in the dimension parallel to the axis. In this dimension the improvement due to the use of multiple beams over other methods is enormous, being comparable with the power of the electron microscope in the other two dimensions.

The details which are revealed are those which introduce a phase difference. At intervals of $\lambda/2$ of optical distance from the surface of the reference flat, plane parallel zones in space may be imagined, of thickness only a few per cent of the distance between zones. If a detail on the surface under investigation is to appear at all in the pattern observed, it must fall within these zones. For the greater part of each zone there is a visible change of intensity between parallel planes distance apart only a few ångströms. Between the narrow zones of high intensity gradient there is practically no light and no intensity gradient, and a step of 2000 A. can escape notice.

When studying a surface, therefore, it is essential that every point of the surface must at some time fall within a zone of maximum light.

Up to this point the conditions to ensure that the final image is similar to the fringe pattern localized in the wedge have been treated. Whether these fringes follow the true contours will now be considered.

In the general case of a surface against an optical flat, a detail is observed either as an area of different tint or as a short displaced part of a fringe. The only occasion when the interpretation of size or contour may be erroneous is when the region is extremely small. The fringe delineates the true contour if the interference conditions are fulfilled over this small region.

For example, the detail observed may be a flat-bottomed channel perpendicular to the fringes and 0·01 mm. wide. The incident plane wave will be diffracted through large angles and so will each of the multiply-reflected waves. If the optical flat is very close to the surface, say five waves (a distance only a quarter of the width of the channel), the diffraction is restricted to the edges, and inside the channel the waves are effectively plane and the interference conditions are fulfilled over most of the width.

An exact calculation of the effect does not appear to be fruitful. Experimentally, many observers in this laboratory agree that the definition and perfection of details revealed is much greater the lower the order of interference. This also depends on the fact that the lateral displacement of the beams down the wedge is more easily kept within the boundaries of the details when very thin gaps are adopted.

One effect of this diffraction at the surface steps is that a local wing is produced at the fringe break on the side of greater gap. It seems likely that the sharp edge of the fringe represents the true contour of the surface for details down to sizes of 0·01mm.; crystalline cleavage zones of this size can be perfectly resolved. Deductions about minute detail are not too reliable, especially when one considers the critical dependence on focus of fringe detail. The fringe reproduced in figure 6 of Part I of this paper is a typical example of this latter effect, the granular structure of the fringe having an appearance dependent entirely on focus.

Many existing conditions for the effective formation and use of multiple-beam Fizeau fringes are favourable: perfection of the surface is desirable over only highly localized regions; many sources of trouble are readily eliminated by the use of a low interference order, the fringes are easily obtained bright and very sharp. The technique is thus a considerable advance on any previous application of the fringes of equal thickness. The measurement of cleavage steps, surface curvature, etc., can be still further improved by adopting fringes of equal chromatic order (Tolansky, 1945) in place of these fringes. Both types of fringe obey the same fundamental formulae, and all the results described above for Fizeau fringes can be adapted to the description of fringes of equal chromatic order.

REFERENCES

Brossel, J., 1947. " Multiple Beam Fizeau Fringes : Part I." *Proc. Phys. Soc.*, **59**, 224.
Fabry, C., 1927. *Les Applications des Interferences Lumineuses* (Paris : Ed. Rev. Opt. théor. instrum.).
Tolansky, S., 1945. *Phil. Mag.*, **36**, 225.

THE DETERMINATION OF THERMAL LAGGING TIMES

By J. C. EVANS,
The National Physical Laboratory

MS. received 15 *September* 1946

ABSTRACT. The problem discussed in this paper arose out of the need for accurately determining the temperature of a barometer under conditions of varying air temperature. A method for computing the thermal lag of simple and complex bodies of cylindrical form, when the ambient temperature is changing steadily, is given and formulae derived for determining the lagging times of such bodies in liquid and in gaseous media. Values calculated from the formulae are compared with results obtained experimentally.

§ 1. INTRODUCTION

The author was led to consider the question discussed in this paper when engaged, some years ago, in investigating the degree of precision attainable in the measurement of pressure by means of the mercury barometer of normal design. The theory which was developed has ever since served as the basis of temperature measurement in barometric work at the National Physical Laboratory and has proved fully satisfactory. For this reason, and because it may be found of use in other applications, it seems desirable to place the theory on record.

The temperature coefficient of a Fortin barometer of normal design is large, a change of temperature of 1° c. corresponding to a change of 0·12 mm. in the reading when the pressure is constant and of the order of 760 mm. of mercury. In precision measurements, pressures are required to 0·01 mm. or better, and a determination of the true temperature of the barometric column is therefore essential. As ordinarily supplied, the barometer is fitted with a thermometer mounted on the brass sheath, but with its bulb between the sheath and the mercury tube. Under conditions of varying air temperature, this arrangement clearly gives only the approximate temperature of the barometric column, and for more specialized applications other methods have been introduced. For example, the thermometer is sometimes mounted with its bulb immersed in the mercury in the cistern of the barometer; alternatively, the thermometer is mounted alongside the barometer, its bulb immersed in mercury held in a glass tube of the same bore as the barometer tube. Experience has shown that in the former case the temperature of the thermometer lags behind that of the barometric column, whilst in the latter the temperature of the column lags behind the thermometer.

§ 2. THEORETICAL BASIS OF CORRECTION

An allowance for the error which results from the use of an incorrectly matched thermometer may be made if the "lagging times" of the column and of the thermometer can be determined; alternatively, the lagging times can be made the same by providing a suitably shielded thermometer. The definition of lagging time emerges from the following considerations.:—

(1) Assume a body at temperature θ_0 immersed at time $t=0$ in a medium at the higher temperature ϕ. The initial rate of change of temperature of the body, assuming Newton's Law, is proportional to the temperature difference $\phi-\theta_0$.

If θ is the temperature at time t,

$$\frac{d\theta}{dt}=\frac{1}{\lambda}(\phi-\theta),$$

in which λ is a constant, and therefore

$$\phi-\theta=(\phi-\theta_0)e^{-\frac{t}{\lambda}}. \qquad \ldots\ldots(1)$$

The constant λ is known as the lagging time of the body in the medium postulated. After the lapse of an interval λ the difference between the temperature of the body and that of the medium will have been reduced to $1/e$ of its initial value.

(2) Assume that a body at temperature θ_0 is immersed in a medium at the same temperature and that at time $t=0$ the temperature of the medium commences to rise uniformly at the rate K per unit time.

Then, at time t,

$$\frac{d\theta}{dt}=\frac{1}{\lambda}(\theta_0+Kt-\theta),$$

which gives $\qquad \theta_0+Kt-K\lambda-\theta=Ae^{-\frac{t}{\lambda}}$

or, since $\qquad \theta=\theta_0 \quad \text{when} \quad t=0,$

$$\theta_0+Kt-K\lambda-\theta=-K\lambda e^{-\frac{t}{\lambda}}.$$

Hence

$$\left.\begin{aligned}\theta&=\theta_0+Kt-K\lambda(1-e^{-\frac{t}{\lambda}})\\ \frac{d\theta}{dt}&=K(1-e^{-\frac{t}{\lambda}}).\end{aligned}\right\} \qquad \ldots\ldots(2)$$

and

After the lapse of a period of time which is large compared with λ, the rate of change of temperature of the body is identical with that of the medium, but the temperature of the body lags behind that of the medium by the amount $K\lambda$, i.e. by the product of the lagging time and the rate of change of temperature.

If two bodies (e.g. a barometer and a thermometer) are immersed in the medium and λ_1 and λ_2 are their lagging times, their temperature difference will be $K(\lambda_1-\lambda_2)$.

§3. LAGGING TIME OF BODY OF CYLINDRICAL FORM IMMERSED IN LIQUID

A. *Theoretical computation*

The basis used for computing the lagging time theoretically is as follows :—

Consider an infinitely long cylindrical glass tube filled with mercury which is immersed in a medium whose temperature is rising at a constant rate. Let r_1 and r_2 be the internal and external radii of the glass tube; k_g, S_g and ρ_g the thermal conductivity, specific heat and density of glass; k_m, S_m and ρ_m the corresponding quantities for mercury; and let

$$\frac{k_g}{\rho_g S_g}=h_g^2,$$

$$\frac{k_m}{\rho_m S_m}=h_m^2.$$

After a certian lapse of time, the rate of flow of heat into the tube will become steady and the temperature at any point will rise at the same rate as the medium. The isothermal surfaces will be cylinders co-axial with the tube, and if θ is the temperature of an isothermal surface within the glass wall and of radius r, the differential equation expressing the flow of heat across this surface is

$$\frac{\partial\theta}{\partial t}=h_g^2\left(\frac{\partial^2\theta}{\partial r^2}+\frac{1}{r}\frac{\partial\theta}{\partial r}\right).$$

But $\dfrac{\partial\theta}{\partial t}$ is constant, so that for this condition

$$\frac{\partial^2\theta}{\partial r^2}+\frac{1}{r}\frac{\partial\theta}{\partial r}=C_g$$

where
$$C_g=\frac{\partial\theta}{\partial t}\Big/h_g^2.$$

The integral of this equation is

$$\theta=\tfrac{1}{4}C_g r^2+A_g\log_e r+B_g,$$

in which A_g and B_g are constants*.

For an isothermal surface in the mercury, θ is given by the corresponding equation

$$\theta=\tfrac{1}{4}C_m r^2+A_m\log_e r+B_m,$$

in which

$$C_m=\frac{\partial\theta}{\partial t}\Big/h_m^2 \quad\text{and}\quad A_m=0 \quad\text{since } \theta\neq\pm\infty \text{ when } r=0.$$

---

* Their values depend, of course, on the magnitude of $\partial\theta/\partial t$.

The equation giving the temperature of an isothermal surface in the mercury can also be obtained as follows. The heat which flows per unit time across an isothermal surface of radius r in the mercury column serves to raise the temperature of the mercury within this surface at the rate $\partial\theta/\partial t$.

Hence,
$$2\pi r k_m \frac{\partial\theta}{\partial r} = \pi r^2 \rho_m S_m \frac{\partial\theta}{\partial t},$$
i.e.
$$\frac{\partial\theta}{\partial r} = \tfrac{1}{2} r C_m,$$
or
$$\theta = \tfrac{1}{4} C_m r^2 + B_m.$$

At any instant, whilst temperatures are changing at a steady rate, let
$$\theta = \theta_2 \quad \text{when} \quad r = r_2,$$
$$\theta = \theta_1 \quad \text{when} \quad r = r_1,$$
$$\theta = \theta_0 \quad \text{when} \quad r = 0,$$
and let λ_r be the lagging time in relation to the external wall of the tube of an isothermal surface of radius r within the mercury.

Then
$$\lambda_r \frac{\partial\theta}{\partial t} = \theta_2 - \theta_r,$$
where
$$\theta_r = \tfrac{1}{4} C_m r^2 + B_m,$$
or, since
$$\theta_1 = \tfrac{1}{4} C_m r_1^2 + B_m,$$
$$\theta_r = \theta_1 - \tfrac{1}{4} C_m (r_1^2 - r^2).$$

Hence
$$\lambda_r \frac{\partial\theta}{\partial t} = (\theta_2 - \theta_1) + (\theta_1 - \theta_r)$$
$$= \tfrac{1}{4} C_g (r_2^2 - r_1^2) + A_g \log_e \frac{r_2}{r_1} + \tfrac{1}{4} C_m (r_1^2 - r^2).$$

But
$$\frac{\partial\theta}{\partial t} = C_g h_g^2 = C_m h_m^2,$$
and, therefore,
$$\lambda_r = \frac{1}{4h_g^2}(r_2^2 - r_1^2) + \frac{A_g}{C_g h_g^2} \log_e \frac{r_2}{r_1} + \frac{1}{4h_m^2}(r_1^2 - r^2).$$

The value of the constant A_g may be obtained by equating the heat flow per unit time across the inner wall of the glass tube to the heat required to raise the temperature of the mercury column at the constant rate $\partial\theta/\partial t$.

Thus,
$$2\pi r_1 k_g \left(\frac{\partial\theta}{\partial r}\right)_{r=r_1} = \pi r_1^2 \rho_m S_m \frac{\partial\theta}{\partial t},$$
i. e.
$$\left(\frac{\partial\theta}{\partial r}\right)_{r=r_1} = \tfrac{1}{2} r_1 \frac{\rho_m S_m}{\rho_g S_g} \cdot \frac{\rho_g S_g}{k_g} \cdot \frac{\partial\theta}{\partial t}$$
$$= \tfrac{1}{2} C_g \frac{\rho_m S_m}{\rho_g S_g} r_1.$$

Hence $$\tfrac{1}{2}C_g r_1 + \frac{A_g}{r_1} = \tfrac{1}{2}C_g \frac{\rho_m S_m}{\rho_g S_g} r_1,$$

which gives $$A_g = \tfrac{1}{2}C_g \left(\frac{\rho_m S_m - \rho_g S_g}{\rho_g S_g}\right) r_1^2.$$

Hence $$\lambda_r = \frac{1}{4h_g^2}(r_2^2 - r_1^2) + \frac{1}{2h_g^2}\left(\frac{\rho_m S_m - \rho_g S_g}{\rho_g S_g}\right) r_1^2 \log_e \frac{r_2}{r_1} + \frac{1}{4h_m^2}(r_1^2 - r^2)$$

or $$\lambda_r = \frac{\rho_g S_g}{4k_g}(r_2^2 - r_1^2) + \frac{\rho_m S_m - \rho_g S_g}{2k_g} r_1^2 \log_e \frac{r_2}{r_1} + \frac{\rho_m S_m}{4k_m}(r_1^2 - r^2). \quad \ldots\ldots(3)$$

This equation gives the lagging time, in relation to the external surface of the tube, of an isothermal surface of radius r within the mercury. Before theory can be compared with practice, it is necessary to determine the mean value of λ over the range $r=0$ to $r=r_1$.

If λ_{r_1} is the lagging time of the inner wall of the glass tube in relation to the external surface, it is clear from the manner of derivation of λ_r (but see also § 4, B) that

$$\lambda_r = \lambda_{r_1} + \frac{\rho_m S_m}{4k_m}(r_1^2 - r^2),$$

so that the average lagging time λ_a is given by

$$\lambda_a = \lambda_{r_1} + \frac{\rho_m S_m}{4k_m} \cdot \frac{\int_{-r_1}^{+r_1} (r_1^2 - r^2)\,dr}{\int_{-r_1}^{+r_1} dr}$$

$$= \lambda_{r_1} + \frac{\rho_m S_m}{6k_m} r_1^2,$$

i.e.

$$\lambda_a = \frac{\rho_g S_g}{4k_g}(r_2^2 - r_1^2) + \frac{\rho_m S_m - \rho_g S_g}{2k_g} r_1^2 \log_e \frac{r_2}{r_1} + \frac{\rho_m S_m}{6k_m} r_1^2. \quad \ldots\ldots(4)$$

B. *Experimental determination*

The validity of this equation has been tested by determining experimentally the lagging time of a glass tube containing mercury and fitted with a thermometer. The bore of the tube used was 22·5 mm., the wall thickness 1·6 mm. and the length of the mercury column 815 mm. The thermometer passed through a small side tube sealed to the main tube at about the mid-point of the mercury column.

The tube was placed in a room whose temperature is maintained constant at about 35° c. (95° F.) and kept there until its temperature had become steady. Immediately outside the door of the room was placed a large iron vessel fitted with glass windows and filled with water at about 17° c. (62° F.). The lagging time was determined by transferring the tube as quickly as possible from the room to the water bath and taking correlated readings of temperature and time with the tube completely immersed in the water, commencing when the tube was lowered into the water and continuing until the temperature of the mercury had fallen to that of the bath.

Figure 1 shows the results obtained in graphical form. One graph is the plot against time of $\bar{\theta}$ and the other the plot against time of $\log_{10}(\bar{\theta}-\phi)$, $\bar{\theta}$ being the mercury temperature and ϕ the water temperature, which is effectively constant. (The bar notation is employed to indicate that the thermometer used gives the average temperature over the cross section of the mercury column). The value obtained for the lagging time was 25·3 seconds; a repetition of the experiment gave 25·5 seconds. When the values (in c.g.s. units) of r_1 and r_2 and of the physical constants k, S and ρ for glass and mercury are substituted in the equation given above for λ_a, the result obtained is 24 seconds.

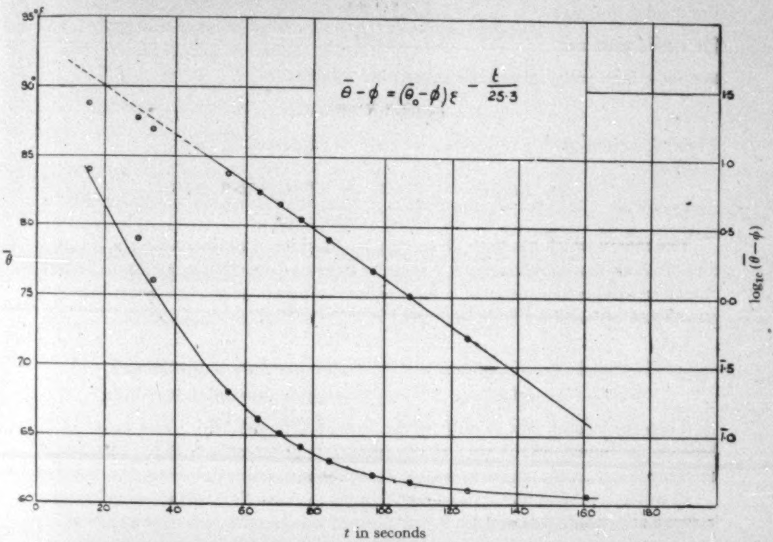


Figure 1. **Observations on mercury column, in glass tube of bore 22·5 mm, immersed in water.**

Attention may be drawn to the fact that the first three observational points plotted in the log graph fall distinctly below the straight line on which the other points lie. This is characteristic of all the curves which have been obtained in experiments on lagging time and is a reminder that the state to which the equations apply is not established instantly. In the experimental verification given above it was, of course, assumed that the outer wall of the glass tube was maintained throughout the observations at the temperature of the water bath; in point of fact, the rise of temperature of the bath during the experiment did not exceed 0°·1 c.

C. *Homogeneous cylinders*

For homogeneous cylinders the equation for λ is simpler. The temperature at a distance r from the axis is given (cf. § 3, A) by

$$\theta=\tfrac{1}{4}Cr^2+B.$$

If θ_1 and θ_0 are the external and axial temperatures and R is the radius of the cylinder

$$\theta_1 - \theta_0 = \tfrac{1}{4}CR^2,$$

so that if λ_0 is the lagging time of the axis in relation to the surface of the cylinder,

$$\lambda_0 Ch^2 = \tfrac{1}{4}CR^2$$

or

$$\lambda_0 = \frac{1}{4h^2}R^2 = \frac{\rho S}{4k}R^2.$$

For a solid iron rod $\lambda_0 = 1{\cdot}7\,R^2$;

for a solid glass rod $\lambda_0 = 44R^2$;

and for a thin-walled glass tube containing toluene

$$\lambda_0 = 220R^2 \text{ approx.}$$

These equations give λ_0 in seconds when R is expressed in cm.

§ 4. LAGGING TIME IN STAGNANT AIR

A. *Theoretical computation*

The theory which has been given may be considered to deal with the lagging time of a body in a liquid medium. In order to extend it so that it applies to bodies immersed in a gaseous medium, it is necessary to introduce the law governing the rate of gain of heat by a body immersed in a gas at higher temperature.

Let

E = emissivity of the surface; θ = temperature of the surface;
A = area of the surface; ϕ = temperature of the medium.

Then the rate of gain of heat by the surface is $EA(\phi - \theta)$. The value of E depends, of course, on the state of the surface (e.g. whether highly polished or blackened) and on the degree of stagnation of the medium.

Consider a tube of liquid immersed in a gaseous medium whose temperature is rising at a steady rate and let

λ_A = lagging time of the axis in relation to the medium.

The lagging time of the axis in relation to the external surface of the tube will be identical with the lagging time in a liquid medium, e.g. water; it will, therefore, be referred to in what follows as the lagging time in water and will be denoted by λ_W. Then if, at any instant,

θ_A = temperature of the gaseous medium,
θ_2 = temperature of external surface of tube,
θ_0 = temperature at the axis, and
K = constant rate of change of temperature of the medium,

it follows that

$$\lambda_W \,.\, K = \theta_2 - \theta_0,$$
$$\lambda_A \,.\, K = \theta_A - \theta_0,$$

whence

$$\lambda_A = \lambda_W + \frac{\theta_A - \theta_2}{K}.$$

The heat gained per unit time by the external wall of the tube is used in raising the temperature of the tube and liquid by an amount K.

Using the same notation as before,

$$2\pi r_2 E(\theta_A - \theta_2) = \{\pi(r_2^2 - r_1^2)\rho_g S_g + \pi r_1^2 \rho_m S_m\}K,$$

so that

$$\frac{\theta_A - \theta_2}{K} = \frac{(r_2^2 - r_1^2)\rho_g S_g + r_1^2 \rho_m S_m}{2r_2 E}$$

Hence

$$\lambda_A = \lambda_W + \frac{(r_2^2 - r_1^2)\rho\, S_g + r_1^2 \rho_m S_m}{2r_2 E}, \qquad \ldots\ldots(5)$$

and is determined since λ_W is given by the earlier equations.

For homogeneous cylinders of radius R,

$$\lambda_A = \lambda_W + \frac{\rho S}{2E} R,$$

i.e.

$$\lambda_A \doteq \frac{\rho S}{4k} R^2 + \frac{\rho S}{2E} R. \qquad \ldots\ldots(6)$$

Some interesting deductions may be made from equation (6). As examples, the lagging times in air of solid rods of glass and of iron will be compared with their lagging times in water. Let the radius of the rods be 1 cm.

The lagging times (to the axis) in water are:—

(1) for the glass rod 44 sec. (cf. § 3, C);
(2) for the iron rod 1·7 sec.

The lagging times in air at normal pressure are:—

(1) for the glass rod 1100 sec.;
(2) for the iron rod 2000 sec.

A value of 0·0002 c.g.s. units* has been adopted for E, the author's experience being that this value appears to be satisfactory for polished surfaces under ordinary atmospheric conditions.

It will be seen that, whereas the lagging time of the iron rod in water is negligible by comparison with that of the glass rod, the lagging time in air is almost twice as great. It is not until the radius of the rods exceeds 20 cm. that the better conductivity of the iron is able to offset the lower thermal capacity per unit volume of the glass. (This argument is of course subject to the assumption that $E = 0{\cdot}0002$ c.g.s. units in both cases.)

If the two 1 cm. rods are immersed in water whose temperature is rising at the rate of 1° c. per hour, the difference between the axial temperature and the temperature of the water is

0°·012 c. for the glass rod;
0°·0005 c. for the iron rod.

On the other hand, if the rods are immersed in air rising in temperature at the same rate, the lag of the axial temperature behind that of the air is

0°·3 c. for the glass rod;
0°·55 c. for the iron rod.

---

* Ingersoll and Zobel, *Mathematical Theory of Heat Conduction*, p. 163 (1913).

B. *Extension to more complex bodies of cylindrical form*

The computation of the lagging time of the barometric column of a barometer is more complex than any of the cases yet considered, since the glass tube is surrounded by a brass sheath. It can, however, be carried out by application of the principles already discussed, although it will simplify matters if two theorems of general application are first stated.

(a) Let 1, 2, 3 .... n be isothermal surfaces in a heterogeneous body immersed in a medium whose temperature is rising at the uniform rate K per unit time.

At any instant, let θ_r and θ_s be the temperatures of the isothermal surfaces r and s and let $\lambda_{r,s}$ be the lagging time of surface r behind surface s.

Then
$$\theta_2-\theta_1=\lambda_{1,2}\,.\,K,$$
$$\theta_3-\theta\;=\lambda_{2,3}\,.\,K,$$
$$\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot$$
$$\theta_n-\theta_{n-1}=\lambda_{n-1,n}\,.\,K.$$
$$\therefore\quad \theta_n-\theta_1=\{\lambda_{1,2}+\lambda_{2,3}+\;\ldots\;+\lambda_{n-1,n}\}\,.\,K.$$
But
$$\theta_n-\theta_1=\lambda_{1,n}\,.\,K,$$
so that
$$\lambda_{1,n}=\lambda_{1,2}+\lambda_{2,3}+\;\ldots\;+\lambda_{n-1,n}. \qquad \ldots\ldots(7)$$
Particular examples of this theorem have already been given.

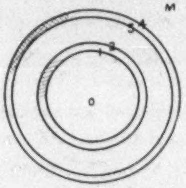


Figure 2. Section of barometer of normal design (not to scale).

(b) Consider a body bounded by a surface of area A in a gaseous medium whose temperature is rising at the rate K per unit time.

Let
λ = lagging time of surface in relation to the medium;
ϕ = temperature of medium at any instant;
θ = temperature of surface at same instant;
U = thermal capacity of the body;
E = emissivity of the surface.

Then
$$\lambda\,.\,K=\phi-\theta$$
and
$$E\,.\,A\,.(\phi-\theta)=U.K.$$
Hence
$$\lambda=\frac{\phi-\theta}{K}=\frac{U}{EA}. \qquad \ldots\ldots(8)$$

Figure 2 shows a section of a barometer of normal design immersed in a gaseous medium M. O represents the axis of the barometer, 1 and 2 the surfaces of the glass tube and 3 and 4 the surfaces of the metal sheath. It is assumed that the glass and metal tubes are co-axial, so that 1, 2, 3 and 4 are isothermal surfaces.

Let r_1 = internal radius of the glass tube;
r_2 = external radius of the glass tube;
r_3 = internal radius of the metal tube;
r_4 = external radius of the metal tube.

Let $\lambda_{4,\mathrm{M}}$ be the lagging time of surface 4 in relation to the medium, $\lambda_{3,4}$ the lagging time of surface 3 in relation to surface 4, etc.

Then the lagging time of the axis behind the medium is given by

$$\lambda_{0,\mathrm{M}} = \lambda_{0,1} + \lambda_{1,2} + \lambda_{2,3} + \lambda_{3,4} + \lambda_{4,\mathrm{M}}.$$

But $$\lambda_{0,1} + \lambda_{1,2} = \lambda_{0,2} = \lambda_{\mathrm{W}},$$

and has already been determined.

$$\lambda_{2,3} = \frac{U}{EA} = \frac{\rho_g S_g (r_2^2 - r_1^2) + \rho_m S_m r_1^2}{2 r_2 E}.$$

$\lambda_{3,4}$ is obtained as follows:—

When the rate of change of temperature has become constant, temperatures in the metal sheath are given (cf. § 3, A) by the equation

$$\theta = \tfrac{1}{4} C_b r^2 + A_b \log_e r + B_b,$$

in which the suffix *b* denotes that the constants are for brass.

The flow of heat per unit time across unit length of surface 3 is

$$2\pi r_3 k_b \left(\frac{\partial \theta}{\partial r}\right)_{r=r_3} = 2\pi r_3 k_b \left(\tfrac{1}{2} C_b r_3 + \frac{A_b}{r_3}\right).$$

The thermal capacity of the medium between surfaces 2 and 3 is negligibly small and it may be assumed that the heat flow from surface 3 is entirely absorbed by the barometer tube.

Hence $$2\pi r_3 k_b \left(\tfrac{1}{2} C_b r_3 + \frac{A_b}{r_3}\right) = 2\pi r_2 E (\theta_3 - \theta_2),$$

where θ_2 and θ_3 are simultaneous values of the temperatures of surfaces 2 and 3.

But $$\theta_3 - \theta_2 = \lambda_{2,3} \cdot \frac{\partial \theta}{\partial t} = \lambda_{2,3} C_b h_b^2,$$

so that $$\tfrac{1}{2} r_3^2 k_b C_b + k_b A_b = r_2 E \lambda_{2,3} C_b h_b^2,$$

i.e. $$A_b = \tfrac{1}{2} C_b \left(\frac{2 r_2 E}{k_b} \lambda_{2,3} h_b^2 - r_3^2\right).$$

Hence

$$\lambda_{3,4} = \frac{\theta_4 - \theta_3}{C_b h_b^2} = \frac{C_b (r_4^2 - r_3^2)}{4 C_b h_b^2} + \frac{C_b \left(\frac{2 r_2 E}{k_b} \lambda_{2,3} h_b^2 - r_3^2\right)}{2 C_b h_b^2} \log_e \frac{r_4}{r_3},$$

i.e. $$\lambda_{3,4} = \frac{\rho_b S_b (r_4^2 - r_3^2)}{4 k_b} - \frac{\rho_b S_b r_3^2}{2 k_b} \log_e \frac{r_4}{r_3} + \frac{r_2 E}{k_b} \log_e \frac{r_4}{r_3} \lambda_{2,3}.$$

It is unnecessary to substitute for $\lambda_{2,3}$ since, for purposes of calculation, $\lambda_{2,3}$ must be evaluated and the value obtained can be introduced when computing $\lambda_{3,4}$.

$$\lambda_{4,\mathrm{M}} = \frac{U}{EA} = \frac{\rho_b S_b (r_4^2 - r_3^2) + \rho_g S_g (r_2^2 - r_1^2) + \rho_m S_m r_1^2}{2 r_4 E}.$$

This may be simplified for computing purposes as follows:—

$$\lambda_{4,\,\mathrm{M}}=\frac{\rho_b S_b(r_4^2-r_3^2)}{2r_4E}+\frac{\rho_g S_g(r_2^2-r_1^2)+\rho_m S_m r_1^2}{2r_2E}\times\frac{r_2}{r_4}$$

$$=\frac{\rho_b S_b(r_4^2-r_3^2)}{2r_4E}+\frac{r_2}{r_4}\lambda_{2,\,3}.$$

Hence, finally,

$$\begin{aligned}\lambda_{0,\,\mathrm{M}}=&\frac{\rho_g S_g(r_2^2-r_1^2)}{4k_g}+\frac{\rho_m S_m-\rho_g S_g}{2k_g}r_1^2\log_e\frac{r_2}{r_1}+\frac{\rho_m S_m r_1^2}{4k_m}\\&+\frac{\rho_g S_g(r_2^2-r_1^2)+\rho_m S_m r_1^2}{2r_2E}\\&+\frac{\rho_b S_b(r_4^2-r_3^2)}{4k_b}-\frac{\rho_b S_b r_3^2}{2k_b}\log_e\frac{r_4}{r_3}+\frac{r_2E}{k_b}\log_e\frac{r_4}{r_3}\lambda_{2,\,3}\\&+\frac{\rho_b S_b(r_4^2-r_3^2)}{2r_4E}+\frac{r_2}{r_4}\lambda_{2,\,3}.\end{aligned}\qquad\ldots\ldots(9)$$

It remains to point out that in order to obtain a mean lagging time for the mercury column instead of the lagging time to the axis, the denominator of the last term in the expression for λ_{W} is changed from $4k_m$ to $6k_m$ (cf. end of § 3, A).

For purposes of calculation, the form given above for $\lambda_{0,\,\mathrm{M}}$ is most suitable; it is of interest, however, to express $\lambda_{0,\,\mathrm{M}}$ in terms collected separately for the mercury, glass and metal components, viz.,

$$\begin{aligned}\lambda_{0,\,\mathrm{M}}=&\rho_m S_m r_1^2\left\{\frac{1}{4k_m}+\frac{1}{2k_g}\log_e\frac{r_2}{r_1}+\frac{1}{2r_2E}+\frac{1}{2k_b}\log_e\frac{r_4}{r_3}+\frac{1}{2r_4E}\right\}\\&+\rho_g S_g(r_2^2-r_1^2)\left\{\frac{1}{4k_g}+\frac{1}{2r_2E}+\frac{1}{2k_b}\log_e\frac{r_4}{r_3}+\frac{1}{2r_4E}\right\}-\frac{\rho_g S_g r_1^2}{2k_g}\log_e\frac{r_2}{r_1}\\&+\rho_b S_b(r_4^2-r_3^2)\left\{\frac{1}{4k_b}+\frac{1}{2r_4E}\right\}-\frac{\rho_b S_b r_3^2}{2k_b}\log_e\frac{r_4}{r_3}.\end{aligned}\qquad\ldots\ldots(10)$$

The following table gives the lagging times of the barometric columns of three barometers of different sizes, but all of ordinary design, as determined by means of equation (9). The first barometer is one of Meteorological Office pattern, the other two are working standard barometers used at the National Physical Laboratory:—

Table 1

Dimensions of barometer	Lagging time in air (minutes)
Internal diameter of glass tube = 8 mm. External diameter of brass tube = 25 mm.	20
Internal diameter of glass tube = 16 mm. External diameter of brass tube = 34 mm.	35
Internal diameter of glass tube = 20 mm. External diameter of brass tube = 39 mm.	45

A value of 0·0002 c.g.s. units for E has again been adopted in making the calculations. It should be pointed out that the brass tube of a barometer of ordinary design is slotted at its upper end and that no allowance for this departure from the theoretical assumptions has been made in the calculations.

C. *Experimental determinations*

Experimental determinations of the lagging times of mercury barometers have given results in reasonably good agreement with the values tabulated above, but the measurements are difficult to make and the best confirmation of the principles discussed has been the substantial improvement in accuracy obtained when the temperature of the barometer is determined by means of an appropriately lagged thermometer, i.e. one which has a lagging time equal to that of the barometer as obtained by calculation. A lagging of calculable effect can be provided by immersing the bulb of the thermometer in a glass tube containing mercury and surrounding the tube with one or more metal sheaths. The lagging time is computed in the manner already described and the following table gives results for typical cases:—

Table 2

Description of lagging	Lagging time in air (minutes)
Unshielded thermometer with thin-walled glass bulb of diameter 5 mm.*	3
Same thermometer shielded with brass sheath of diameter 20 mm., wall thickness 1 mm.	10
Bulb of thermometer immersed in mercury contained in glass tube of bore 16 mm. and wall thickness 2·5 mm.	20
Thermometer in same tube with iron sheath of internal diameter 30 mm. and wall thickness 4 mm.	52
Thermometer in same tube with iron sheath as above and an outer brass sheath of internal diameter 44 mm. and wall thickness 4 mm.	95

* The lagging time of this thermometer in water is 3 to 4 seconds.

These values also are based on a value of 0·0002 c.g.s. units for E.

It is of interest to show the contributions of the various terms λ_W, $\lambda_{2,3}$, etc. to the total lagging time $\lambda_{0,M}$. For the thermometer with the single iron sheath cited in the table above,

$$\lambda_W = 30 \text{ sec.} \qquad \lambda_{2,3} = 1100 \text{ sec.}$$

$$\lambda_{3,4} \text{ is quite negligible} \quad \text{and} \quad \lambda_{4,M} = 2000 \text{ sec.}$$

The lagging time between the surfaces of the metal sheath can then, for all practical purposes, be neglected and this leads to a considerable simplification of the calculations for cases such as the last referred to in table 2. The lagging time for this thermometer is

$$\lambda_{0,M}=\lambda_{0,1}+\lambda_{1,2}+\lambda_{2,3}+\lambda_{3,4}+\lambda_{4,5}+\lambda_{5,6}+\lambda_{6,M},$$

$$\lambda_{0,1}+\lambda_{1,2}=\lambda_W;$$

$\lambda_{3,4}$ and $\lambda_{5,6}$ can be neglected.

By analogy with equation (9) the expressions for the remaining terms are

$$\lambda_{2,3}=\frac{\rho_g S_g(r_2^2-r_1^2)+\rho_m S_m r_1^2}{2r_2E},$$

$$\lambda_{4,5}=\frac{\rho_f S_f(r_4^2-r_3^2)}{2r_4E}+\frac{r_2}{r_4}\lambda_{2,3},$$

$$\lambda_{6,M}=\frac{\rho_b S_b(r_6^2-r_5^2)}{2r_6E}+\frac{r_4}{r_6}\lambda_{4,5}.$$

Practical determinations of the lagging times in air of thermometers immersed in mercury and surrounded by one or two metal sheaths (dimensions of glass and metal tubes as given in table 2) have been made. The experimental procedure was similar to that already described in relation to the mercury tube. The iron chamber was used again, but this time with a view to obtaining reasonably stagnant air conditions. The thermometers (one with a single sheath, the other with two) were mounted on a slab of cork 2 inches thick, the stems of the thermometers being 3 inches apart. Annular grooves of small depth were cut in the cork to locate the mercury tubes and the metal sheaths, a little plasticene being used to provide grip. The cork slab was suspended by means of brass strips from a wooden support shaped so as to form a lid for the iron chamber. For purposes of comparison an unshielded thermometer was suspended between the other two.

Experiments were made under both falling and rising conditions of temperature. Several hours before commencing the test under falling conditions, the thermometers were placed in a room maintained at 35° c. approximately and the iron chamber was placed immediately outside the door of the room. For the other experiment, the iron chamber was placed in the hot room whilst the thermometers were kept in an adjoining room which was maintained at 20° c.

The actual thermometric readings obtained in the experiments are not given here but the logarithmic plots are shown in figures 3 and 4. The results obtained were:—

Thermometer with one sheath

Falling temperature	49 minutes
Rising temperature	50 minutes

Thermometer with two sheaths

Falling temperature	77 minutes
Rising temperature	78 minutes

The lagging time of the unshielded thermometer was, very closely, 3 minutes.

The calculated values for the two shielded thermometers are as given in table 2, viz. 52 and 95 minutes. It must be mentioned that the sheaths and mercury tubes

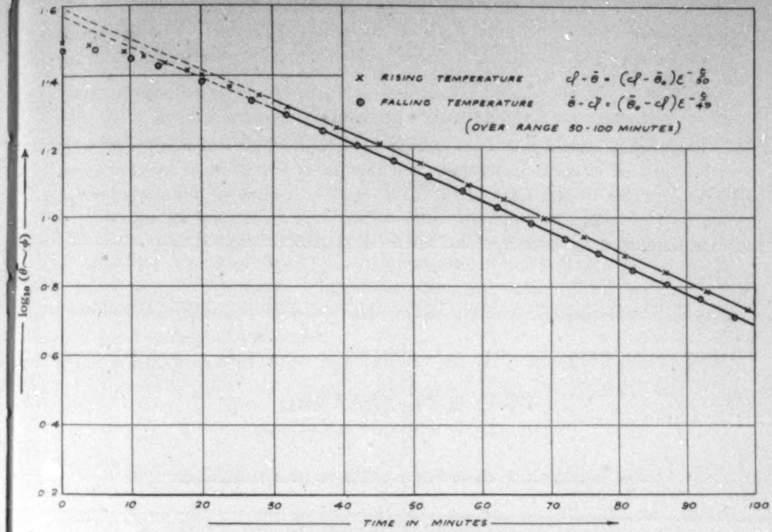


Figure 3. Observations on mercury-immersed thermometer, with single metal sheath, in stagnant air.

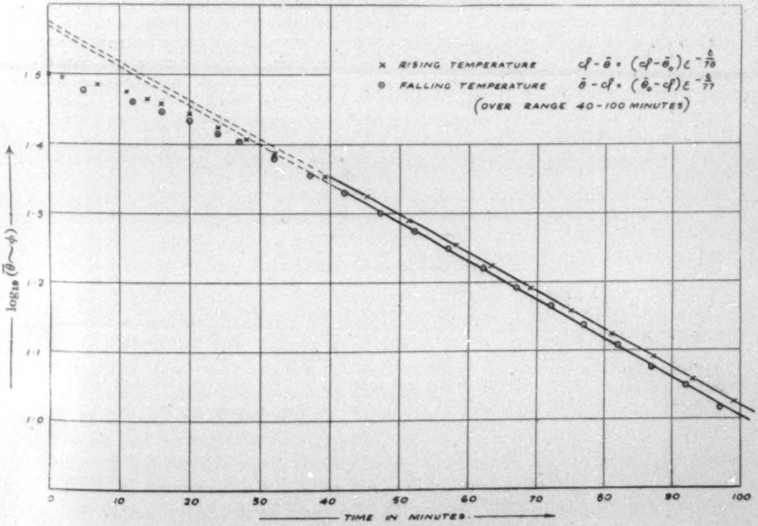


Figure 4. Observations on mercury-immersed thermometer, with double metal sheath, in stagnant air.

used were only about 4 inches long, so that the end effects, particularly for the thermometer with two sheaths, were by no means negligible. The agreement between theory and practice is, however, reasonably good, bearing in mind the uncertainty attaching to the value of E.

§5. ACKNOWLEDGMENTS

The work described above was carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory. The author desires to acknowledge the comments and suggestions he received from Mr. F. A. Gould and the assistance in the experimental work rendered by Mr. F. D. Jones.

DISLOCATIONS IN A SIMPLE CUBIC LATTICE

By F. R. N. NABARRO,
Royal Society Warren Research Fellow

Communicated by N. F. Mott, F.R.S. ; *MS. received* 23 *September* 1946

ABSTRACT. The properties of dislocations are calculated by an approximate method due to Peierls. The width of a dislocation is small, displacements comparable with the interatomic distance being confined to a few atoms. The shear stress required to move a dislocation in an otherwise perfect lattice is of the order of a thousandth of the "theoretical" shear strength. The energy and effective mass of a single dislocation increase logarithmically with the size of the specimen. A pair of dislocations of opposite sign in the same glide plane cannot be in stable equilibrium unless they are separated by a distance of the order of 10 000 lattice spacings. If an external shear stress is applied there is a critical separation of the pair of dislocations at which they are in unstable equilibrium. The energy of this unstable state is the activation energy for the formation of a pair of dislocations. It depends on the external shear, and for practical stresses is of the order of 7 electron volts per atomic plane.

The size and energy of dislocations in real crystals are unlikely to differ greatly from those calculated: the stress required to move a dislocation and the critical separation of two dislocations may be seriously in error.

§1. INTRODUCTION

VOLTERRA's theory of elastic dislocations has been made the basis of a theory of the plastic deformation of crystals. This theory has explained many of the characteristics of plastic deformation, but Volterra's classical assumptions of a continuous medium obeying Hooke's Law even under large strains do not allow a detailed investigation of the displacement of the atoms in the core of a dislocation in a real crystal. Peierls (1940) has shown that by using suitable approximations it is possible to take account of the periodic structure of the crystal in the glide plane, and to determine the size of a dislocation in a simple cubic lattice and the shear stress necessary to move such a dislocation across its glide plane. The first part of this paper amplifies the previous arguments and corrects certain errors in Peierls's paper. In the second part, the solution for a single dislocation is extended to the case of a pair of dislocations of opposite sign held in equilibrium in the same glide plane by an external shear stress. The energy of this system, which is the activation energy for the formation of a pair of dislocations in a stressed crystal, is determined. Finally, a brief discussion is given of the relation of the properties of this idealized model to those of a real crystal.

§2. THE MODEL

The model considered is a simple cubic lattice containing a dislocation of the kind which Burgers (1940) calls "a dislocation of edge type". It is shown in figure 1. The slip plane $P(z=0)$ divides the crystal into an upper part a and a lower part b. These are symmetrical about the vertical plane $S(x=0)$. The central plane S lies in a lattice plane in the upper half-crystal a, and half way between two lattice planes in b.

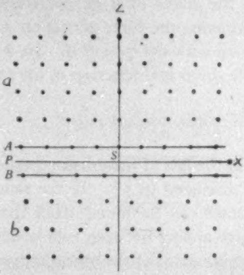


Figure 1. A single dislocation in an unstrained crystal.

The y axis at right angles to the plane of the figure (i.e. the line in which the plane S meets the slip plane P) is called the *dislocation line*. In the neighbourhood of this line the atoms of a are moved inwards and those of b are moved apart, so that at great distances from the dislocation line the two planes A and B, which are immediately above and below the slip plane, are again in correct alignment, A containing, however, one row of atoms more than B. This type of dislocation, which may be formed by cutting a perfect crystal along the half-plane $x=0$, $z>0$, and introducing a single extra layer of atoms into this half-plane, can also be represented as a Taylor dislocation, formed by cutting the same crystal along the half-plane $z=0$, $x<0$ and sliding the upper layer of atoms A to the right over the lower layer B until the relative displacement of the planes at large distances from the dislocation line is one lattice spacing (Taylor, 1934). In Burgers' form, the crystal containing a dislocation is in equilibrium with no external forces, whereas external forces are required to maintain the equilibrium of a crystal containing a dislocation of Taylor's form.

The atoms in the plane A are then subject to two forces : (i) the interaction with other atoms in the half-crystal a; this force tends to spread the compression uniformly over the plane A, i.e. to extend the dislocation: (ii) the interaction with atoms in *b*, particularly with those in the adjacent layer *B*; this has a tendency to bring as many atoms of *A* as possible into correct alignment with *B*, i.e. to shorten the dislocation. In equilibrium these two forces balance.

If the extension of the dislocation is large compared with the atomic distance d, both the horizontal displacement u and the vertical displacement w of the atoms in *A* vary slowly from atom to atom. The relative displacement of neighbouring atoms within each half-crystal is then much smaller than d. In these circumstances each half-crystal may be considered as an elastic continuum. Moreover, for the sake of simplicity, it will be assumed to be elastically isotropic. Then the force (i) is simply the force that has to be applied to the plane surface of an elastic continuum in order to make its horizontal displacement at the surface equal to $u(x)$. This force can be obtained by the usual methods of the theory of elasticity.

To the same degree of approximation, it may be assumed that the horizontal component of the force (ii) depends only on the horizontal displacement of the atoms of *A* relative to the atoms of *B* immediately underneath. If $u(x)$ and $\bar{u}(x)$ are the two displacements, the force acting on a surface element near x is a periodic function of $u-\bar{u}$ with the period d. In a first approximation it may be represented as a simple sinusoidal function of the form

$$\text{const.}\ \sin 2\pi(u-\bar{u})/d. \qquad \ldots\ldots(1)$$

The probable deviation of the law of force from this form, and the consequences of such a deviation, are discussed in §7. If the simple sine law is a sufficient approximation, the constant can be found from the shear modulus, provided it is assumed that the force arising between two lattice planes in a state of shear is independent of the displacement of the other lattice planes. This assumption is not strictly correct, but it is probably a reasonable approximation. By considering a small shear of angle $(u-\bar{u})/d$ it may be seen that the constant is $-\mu/2\pi$.

The dislocation is considered to be very long in the direction of the y axis, and the problem is accordingly one of plane strain. Further, it is assumed that the vertical component of the force (ii) depends only on the relative vertical displacement in *A* with respect to that in *B*. Since the tangential forces applied to *A* and *B* are equal and opposite, their vertical displacements are equal, and the vertical force (ii) vanishes. This assumption again is not actually correct, since in places where the atoms in the two planes are out of alignment, the equilibrium value of their vertical distance will obviously be changed. A separate discussion is required in order to take this effect into account, and it depends on the detailed structure of the crystal lattice.

§3. THE GOVERNING EQUATION

It is first necessary to find the integral equation connecting the displacement $u(x')$ of a point x' on the surface *A* with the tangential stress $p_{xz}(x)$ applied to this surface, assuming that the strains e_{xy}, e_{yy}, e_{zy} vanish across the surface *A* in accordance with the assumption of §2.

The problem may be solved by assuming that the stresses are expressed in terms of a stress function χ in the usual form:

$$\left.\begin{aligned} p_{xx} &= \frac{\partial^2\chi}{\partial z^2}, \\ p_{xz} &= -\frac{\partial^2\chi}{\partial x \partial z}, \\ p_{zz} &= \frac{\partial^2\chi}{\partial x^2}, \end{aligned}\right\} \qquad \ldots\ldots(2)$$

while for plane strain the components of strain can be expressed in terms of the components of stress in the form

$$4\mu(\lambda+\mu)e_{xx} = (\lambda+2\mu)p_{xx} - \lambda p_{zz}, \text{ etc.,}$$

where λ and μ are Lamé's elastic contants. Expressing λ in terms of μ and Poisson's ratio σ by means of the relation

$$\frac{\lambda}{\mu} = \frac{2\sigma}{1-2\sigma}$$

leads to

$$2\mu e_{xx} = (1-\sigma)p_{xx} - \sigma p_{zz}, \qquad \ldots\ldots(3)$$

The integral equation may be obtained by choosing a stress function

$$\chi = Z(e^{-mZ}\cos mx - 1), \qquad \ldots\ldots(4)$$

where $Z = z - \frac{1}{2}d$.

This stress function is easily shown to satisfy the equation $\nabla^4\chi = 0$, and leads to

$$\begin{aligned} p_{xx} &= -m(2-mZ)e^{-mZ}\cos mx, \\ p_{xz} &= m(1-mZ)e^{-mZ}\sin mx, \\ p_{zz} &= -m^2 Z e^{-mZ}\cos mx. \end{aligned}$$

On the surface A, $Z=0$, and p_{zz} vanishes across this surface, as has been assumed. The tangential stress is given by

$$p_{xz} = m\sin mx, \qquad \ldots\ldots(5)$$

while

$$p_{xx} = -2m\cos mx.$$

It follows from (3) that

$$\mu e_{xx} = -(1-\sigma)m\cos mx$$

and by integration that

$$\mu u(x) = -(1-\sigma)\sin mx. \qquad \ldots\ldots(6)$$

A more general displacement $u(x)$ may now be expressed as a Fourier integral of the form

$$u(x) = \frac{1}{\pi}\int_0^\infty \int_{-\infty}^\infty \cos m(x'-x) \,.\, u(x')dx'dm. \qquad \ldots\ldots(7)$$

Since the equations of elasticity are linear, the components of displacement corresponding to each value of m may be considered separately. A typical component is of the form

$$\begin{aligned}&\frac{1}{\pi}\int_{-\infty}^{\infty} \cos m(x'-x) \,.\, u(x')dx' \\ &=\frac{1}{\pi}\int_{-\infty}^{\infty} \cos mx' \,.\, u(x')dx' \,.\, \cos mx \\ &+\frac{1}{\pi}\int_{-\infty}^{\infty} \sin mx' \,.\, u(x')dx' \,.\, \sin mx.\end{aligned}$$

Each term is of the same form as (6) or the corresponding expression involving $\cos mx$, and the corresponding tangential stresses are, therefore, given by (5) and the corresponding expression. The resultant tangential stress is given formally by

$$p_{xz}=-\frac{\mu}{\pi(1-\sigma)}\int_0^{\infty}\int_{-\infty}^{\infty} m\cos m(x'-x) \,.\, u(x')dx'dm. \qquad \ldots\ldots(8)$$

Integrating by parts, and presupposing that $u(\infty)=u(-\infty)=0$, gives

$$\pi(1-\sigma)p_{xz}/\mu=\int_0^{\infty}\int_{-\infty}^{\infty} \sin m(x'-x)\frac{du}{dx'}dx'dm.$$

A second formal integration by parts leads to

$$\pi(1-\sigma)p_{xz}/\mu=\lim_{m\to\infty}\int_{-\infty}^{\infty}\frac{1-\cos m(x'-x)}{x'-x}\frac{du}{dx'}dx', \qquad \ldots\ldots(9)$$

and if du/dx' is a sufficiently regular function, the term involving $\cos m(x'-x)$ tends to 0 as m tends to ∞. (For a rigorous discussion of this type of transformation see Titchmarsh, 1937.) The general relation now becomes

$$p_{xz}=\frac{\mu}{\pi(1-\sigma)}\int_{-\infty}^{\infty}\frac{1}{x'-x}\frac{du}{dx'}dx', \qquad \ldots\ldots(10)$$

where the Cauchy principal value of the integral is taken.

This equation may be compared with that derived from equation (1), which is

$$p_{xz}=-\frac{\mu}{2\pi}\sin\left[2\pi(u-\bar{u})/d\right].$$

Writing $\phi=u-\bar{u}=2u$, elimination of p_{xz} yields

$$\int_{-\infty}^{\infty}\frac{1}{x-x'}\frac{d\phi}{dx'}dx'=(1-\sigma)\sin\frac{2\pi\phi}{d}. \qquad \ldots\ldots(11)$$

This equation differs slightly from that previously given.

§ 4. THE SOLUTION FOR A SINGLE DISLOCATION

The solution which represents a single dislocation with its centre at the origin as in figure 1, is

$$\frac{\phi}{d}=-\frac{1}{\pi}\tan^{-1}\frac{2x(1-\sigma)}{d}. \qquad \ldots\ldots(12)$$

This solution was given by Peierls. It may be verified by substituting in (11), resolving the integral into partial fractions, and taking the Cauchy principal value of the divergent term.

The width of the dislocation is small, the displacement falling to half its extreme value at a distance $d/2(1-\sigma)$ from the central plane S. The original assumption that the dislocation is spread over a large number of atoms has thus led to a contradiction and, if the sinusoidal law of force (1) is valid and the neglect of vertical forces is justified, a dislocation must be confined to a region of linear dimensions only a few atomic radii.

The components of displacement in parts of the crystal away from the slip plane may be obtained by expressing (12) as a Fourier integral in the form

$$u=\tfrac{1}{2}\phi=-\frac{d}{2\pi}\int_0^\infty e^{-\frac{md}{2(1-\sigma)}}\frac{\sin mx}{m}\,dm. \qquad \ldots\ldots(13)$$

Comparing (13) with (6) and (4) shows that the stress function is

$$\chi=-\frac{\mu d}{2\pi(1-\sigma)}\int_0^\infty e^{-\frac{md}{2(1-\sigma)}}Z\,\frac{e^{-mZ}\cos mx-1}{m}\,dm. \qquad \ldots\ldots(14)$$

The displacements corresponding to (4) are given (Love, 1944) by

$$\left.\begin{aligned}2\mu u&=[mZ-2(1-\sigma)]e^{-mZ}\sin mx,\\ 2\mu w&=[(mZ-1)+2(1-\sigma)]e^{-mZ}\cos mx+1-2(1-\sigma),\end{aligned}\right\} \qquad \ldots\ldots(15)$$

and the displacements corresponding to (14) follow immediately. The singularity at the origin, which is represented in elastic dislocation theory by a factor $[x^2+z^2]$ in the denominator, now disappears, and this factor is replaced by $[x^2+(z+\zeta)^2]$, where $\zeta=d/2(1-\sigma)$.

The energy of unit length of such a dislocation in an infinite crystal is infinite, for the shearing strain at a large distance r from the axis is of the form B/r. The energy density is of the order $1/r^2$, and the total elastic energy of order $\int^\infty(1/r^2)2\pi r\,dr$, which diverges. A finite energy is obtained only by considering a pair of dislocations of opposite sign.

A dislocation has an effective mass, for as it moves across the glide plane the displacements of all points in the body alter. If the dislocation moves across the glide plane with velocity V, the components of velocity of any other point in the body are $V\,\partial u/\partial x$, $V\,\partial w/\partial x$. If the density of the crystal is ρ, the kinetic energy associated with unit length of dislocation is

$$\tfrac{1}{2}\rho V^2\iint\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial w}{\partial x}\right)^2\right]dx\,dz.$$

This is of the same form as the elastic energy, and diverges logarithmically.

If the dislocation is situated in a finite crystal of linear dimensions L, the elastic energy of unit length of the dislocation is of order $\mu d^2\log L/d$, and its effective mass per unit length is of order $\rho d^2\log L/d$.

§5. THE SOLUTION FOR A PAIR OF DISLOCATIONS

It has been suggested that plastic flow may be initiated by the simultaneous production of a pair of dislocations of opposite sign moving in the same glide plane. The form of a crystal under external shear stress, and containing such a pair of dislocations, is shown in figure 2. It is similar to the *Verhakung* discussed by Dehlinger (1929). Such a system cannot be in stable equilibrium. The external stress causes the dislocations to separate, and their separation leads to a shearing motion of the upper and lower parts of the crystal which yields to the external stress. The stress field surrounding each dislocation opposes the external stress in the neighbourhood of the other dislocation, and gives rise to an attraction between the dislocations. For a given applied stress T there is a critical separation $r(T)$ at which the dislocations are in unstable equilibrium, and it is this state of unstable equilibrium which is calculated.

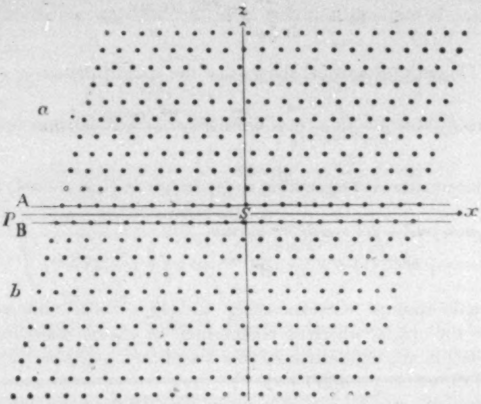


Figure 2. A pair of dislocations in a sheared crystal. The angle of shear shown is ten times as great as that required to maintain equilibrium in the planes A and B.

In the case of figure 1 it was convenient to measure u and $\bar{u}$ from positions such that $\phi=0$ represented the greatest possible misfit of the lattices above and below the glide plane: for the double dislocation of figure 2 it is more convenient to represent this misfit by $\phi=\frac{1}{2}d$ and exact registration of the lattices by $\phi=0$ or $\phi=d$. This changes the sign of the sine function in equations (1) and (11).

The solution for two dislocations is then

$$\frac{\phi}{d}=\frac{1}{\pi}\cot^{-1}\left[\frac{(1-\sigma)^2\sin\theta}{d^2}x^2-\cot\theta\right]+\frac{\theta}{4\pi}, \qquad \ldots\ldots(16)$$

where θ is a parameter defining the separation of the dislocations and that value of $\cot^{-1}$ is taken which lies between 0 and π. This solution may be verified by substitution.

The displacements in rows A and B in figure 2 correspond to $\theta=\pi/10$.

Substituting in (10), the tangential stress required to maintain this displacement is

$$p_{xz}=\frac{\mu}{2\pi}\left(\sin\frac{2\pi\phi}{d}-\sin\frac{\theta}{2}\right), \qquad \ldots\ldots(17)$$

whereas the force produced by the atoms in the other half of the crystal is

$$p_{xz}=\frac{\mu}{2\pi}\sin\frac{2\pi\phi}{d}. \qquad \ldots\ldots(18)$$

The actual tangential stress (18) exceeds the stress (17) required to maintain the displacement (16) by a constant quantity $(\mu/2\pi)\sin\frac{1}{2}\theta$. The upper and lower halves of the crystal can only remain in equilibrium if external stresses of this amount are applied to them. If θ is small, this stress, $\mu\theta/4\pi$, produces a uniform shearing strain of $\theta/4\pi$, which agrees with the strain between rows A and B given by (16) for large values of x.

If θ is small the two dislocations are widely separated. Their centres may be defined as the points $x=\pm d(\cos\theta)^{\frac{1}{2}}/(1-\sigma)\sin\theta$, or for small θ, $x=\pm d/(1-\sigma)\theta$.

It is now possible to calculate the energy of this pair of dislocations. Since the dislocations are embedded in an infinite crystal under a uniform shear stress $\mu\theta/4\pi$, the total energy of the system is infinite. The energy which it is important to calculate is the activation energy required to produce the equilibrium configuration of figure 2 when the crystal is already uniformly stressed. Once this energy is supplied, further separation of the dislocations releases energy.

The energy of the crystal in figure 2 differs from the energy of the uniformly strained crystal by three contributions:

(a) the forces acting across the boundary P do work on each half-crystal;

(b) the potential energy of attraction between the atoms in rows A and B is increased;

(c) the separation of the dislocations represents a shear of one half-crystal over the other, and in this motion work is done by the external shearing forces.

These contributions may be evaluated separately.

(a) *Work done by forces acting across P*

Each half of the crystal is assumed to obey Hooke's Law, and its elastic energy may be expressed in terms of the work done by the forces acting across the plane P by integrals of the form $\frac{1}{2}\int u_x p_{xz}dx$. If the energy is referred to the crystal under uniform shear as zero, the total energy is

$$\begin{aligned}-2\int_{-\infty}^{\infty}\tfrac{1}{2}[u(x)p_{xz}(x)-u(\infty)p_{xz}(\infty)]\,dx\\ =-\frac{\mu}{4\pi}\int_{-\infty}^{\infty}\left[\phi\sin\frac{2\pi\phi}{d}-\frac{d\theta}{4\pi}\sin\frac{\theta}{2}\right]dx,\end{aligned} \qquad \ldots\ldots(19)$$

where ϕ is given by (16).

The expression (19) may be evaluated by writing $x = d(\cos\theta)^{\frac{1}{2}}y/(1-\sigma)\sin\theta$, and becomes

$$-\frac{\mu d^2(\cos\theta)^{\frac{1}{2}}}{4\pi^2(1-\sigma)\sin\theta}\int_{-\infty}^{\infty}[\chi\sin 2\chi - \tfrac{1}{4}\theta\sin\tfrac{1}{2}\theta]\,dy, \qquad \ldots\ldots\ 20)$$

where
$$\cot(\chi - \tfrac{1}{4}\theta) = (y^2-1)\cot\theta.$$

It is easily shown that the integral in (20) is equal to the integral of

$$\frac{1}{2i}\cdot\frac{2(y^2-1)\cot\theta\cos\frac{1}{2}\theta + [(y^2-1)^2\cot^2\theta - 1]\sin\frac{1}{2}\theta}{(y^2-1)^2\cot^2\theta+1}\log\frac{(y^2-1)\cot\theta+i}{(y^2-1)\cot\theta-i}$$
$$+\frac{\theta}{2}\cdot\frac{(y^2-1)\cot\theta\cos\frac{1}{2}\theta - \sin\frac{1}{2}\theta}{(y^2-1)^2\cot^2\theta+1}$$

taken anti-clockwise round the contour of figure 3. This function has essential singularities of the form $(\log z)/z$ at the points P, $Q = \pm(\cos\theta)^{-\frac{1}{2}}e^{\pm\frac{1}{2}i\theta}$. If cuts are made joining P and Q to the point $i\eta$, where η is real and very small, and corresponding cuts in the lower half-plane, the function is single-valued and the logarithm changes by $\pm i\pi$ on crossing the boundaries OP, OQ. The term independent of the logarithm is always single-valued, and has only simple poles at P and Q. The residues are equal and opposite, and this term makes no contribution to the integral. The contribution of the other may be evaluated by taking the loops round P and Q to be equal circles of small radius ϵ. The sum of the integrals round these circles is

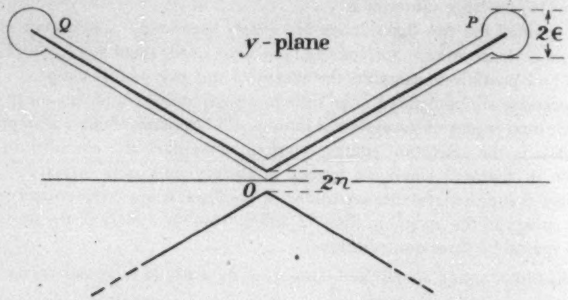


Figure 3. Cuts and contours for the integration of (20).

$$\frac{\pi\sin\theta}{(\cos\theta)^{\frac{1}{2}}}\left[\log\frac{\sin\theta}{(\cos\theta)^{\frac{1}{2}}} - \log\epsilon\right]. \qquad \ldots\ldots(21)$$

On the straight contours OP, PO write $y = \rho(\cos\theta)^{-\frac{1}{2}}e^{\frac{1}{2}i\theta}$, and the limits of ρ are 0 and $1-\epsilon(\cos\theta)^{\frac{1}{2}}$.

The integrand on OP exceeds the integrand at the neighbouring point on PO by

$$\xi(\rho) = \pi\frac{2[\cot\theta + i\rho^2/(\rho^2-1)]\cos\frac{1}{2}\theta + [(\rho^2-1)\cot^2\theta + 2i\rho^2\cot\theta - (\rho^4+1)/(\rho^2-1)]\sin\frac{1}{2}\theta}{(\rho^2-1)\cot^2\theta + 2i\rho^2\cot\theta - (\rho^2+1)}$$

and the sum of the integrals along these two contours OP, PO is

$$\frac{e^{\frac{1}{2}i\theta}}{(\cos\theta)^{\frac{1}{2}}}\int_0^{1-\epsilon(\cos\theta)^{\frac{1}{2}}} \xi(\rho)d\rho.$$

The sum of the integrals along OQ and QO is the complex conjugate of this. In the limit of small ϵ the sum of these four integrals is $\pi\sin\theta(\cos\theta)^{-\frac{1}{2}}$ times

$$1+\log\tan\tfrac{1}{2}\theta+\log\epsilon+\tfrac{1}{2}\log\cos\theta-\log 2. \qquad \ldots\ldots(22)$$

On adding to this the contribution of the small circular contours, the terms in $\log\epsilon$ cancel, and the integral in (20) becomes $\pi\sin\theta(\cos\theta)^{-\frac{1}{2}}F$, where

$$F=1+\log\tan\tfrac{1}{2}\theta+\log\sin\theta-\log 2. \qquad \ldots\ldots(23)$$

The contribution to the total energy of the forces acting across the plane P is now

$$-\mu d^2F/4\pi(1-\sigma). \qquad \ldots\ldots(24)$$

(b) *Potential energy of attraction between A and B*

The potential energy, referred to the homogeneously strained crystal, is

$$\frac{\mu d^2}{4\pi^2}\int_{-\infty}^{\infty}[\cos\tfrac{1}{2}\theta-\cos 2\pi\phi/d]\,dx. \qquad \ldots\ldots(25)$$

On substituting for ϕ and again changing the variable, as in equations (19) and (20), this becomes

$$\frac{\mu d^2(\cos\theta)^{\frac{1}{2}}}{4\pi^2(1-\sigma)\sin\theta}\int_{-\infty}^{\infty}[\cos\tfrac{1}{2}\theta-\cos 2\chi]\,dy. \qquad \ldots\ldots(26)$$

The integrand has simple poles in the upper half-plane at $y=\pm(\cos\theta)^{-\frac{1}{2}}e^{\pm\frac{1}{2}i\theta}$. On evaluating the residues the integral is easily shown to be $2\pi\sin\theta/(\cos\theta)^{\frac{1}{2}}$, and the contribution of the attraction between A and B to the energy is

$$\mu d^2/2\pi(1-\sigma). \qquad \ldots\ldots(27)$$

As was to be expected, this contribution remains finite when θ tends to zero.

(c) *Work done by external forces*

If a uniform shear stress $(\mu/2\pi)\sin\frac{1}{2}\theta$ is maintained on the external surface of the crystal during the formation of the pair of dislocations, the external forces do work

$$\frac{\mu}{2\pi}\sin\tfrac{1}{2}\theta\int_{-\infty}^{\infty}\left[\phi(x)-\frac{d\theta}{4\pi}\right]dx. \qquad \ldots\ldots(28)$$

As before, this may be written

$$\frac{\mu d^2\sin\frac{1}{2}\theta(\cos\theta)^{\frac{1}{2}}}{2\pi^2(1-\sigma)\sin\theta}\int_{-\infty}^{\infty}\cot^{-1}[(y^2-1)\cot\theta]\,dy. \qquad \ldots\ldots(29)$$

Integrating by parts, the integral in (29) becomes

$$2\tan\theta\int_{-\infty}^{\infty}y^2\,dy/[(y^2-1)^2+\tan^2\theta],$$

which may be evaluated from its residues as $2\pi\cos\frac{1}{2}\theta/(\cos\theta)^{\frac{1}{2}}$. The work done by the external forces, which is to be subtracted from the internal energy of the crystal in order to obtain the activation energy for the formation of a pair of dislocations, is, therefore,

$$\mu d^2/2\pi(1-\sigma). \qquad \ldots\ldots(30)$$

This is finite for small θ, because it represents the work done by forces of order θ acting through distances of order $1/\theta$.

The total activation energy is (24)+(27)−(30). For small θ this is approximately $\mu d^2 \log(2/\theta)/2\pi(1-\sigma)$. Its dependence on the distance between the dislocations agrees with that obtained by Koehler (1941).

To estimate the order of magnitude of this quantity, take $\mu = 4{\cdot}4 \times 10^{11}$ dynes/cm², $\sigma = \frac{1}{3}$, and for d take the distance of closest approach of two copper atoms, $2{\cdot}5 \times 10^{-8}$ cm. Then the coefficient $\mu d^2/2\pi(1-\sigma)$ is 6×10^{-5} erg/cm., or $1{\cdot}5 \times 10^{-12}$ erg per atom-pair. This is 1 electron volt for each atomic plane even for large strains of the order $\theta = 1$. For the practical elastic limits of metallic single crystals $\theta \simeq 10^{-3}$ and the energy is 7 electron volts per atomic plane.

§6. THE SHEAR STRESS REQUIRED TO MOVE A SINGLE DISLOCATION

In the previous section, the energy of a pair of dislocations has been calculated by treating the two halves of the crystal as elastic continua, and integrating over the interfaces A and B. To this approximation the energy of a single dislocation in a crystal free from external stress is independent of the position of the dislocation in the crystal. The dislocation is in neutral equilibrium, and will move under the smallest external shearing stress, since its motion causes a relative displacement of the half-crystals a and b in the x-direction. On the other hand if the atomic structure is taken into account, the energy of a dislocation must, in the absence of a stress, depend on its exact position, i.e. on whether the plane of symmetry passes through a row of atoms or not. Hence the dislocation will have a number of positions of stable equilibrium, and these will persist even under a stress until this exceeds a certain magnitude.

The stress required to move a dislocation may be estimated by assuming that each half-crystal a and b retains its form as the dislocation moves, so that the contribution analogous to (a) in §5 is independent of the position of the dislocation. The contribution analogous to (b) is to be evaluated by replacing the integral (25) by the corresponding sum over all atoms in the planes A and B. (It is not sufficient to sum over the atoms of plane A alone, for this would yield an interaction energy which was a periodic function of the displacement of the dislocation with period d. The geometrical form of the dislocation, and therefore its energy, is in fact restored by a displacement of only $\frac{1}{2}d$.) Since the energy is an energy of interaction between the atoms of A and the atoms of B, the sum must be halved to obtain the energy. Taking αd as the displacement of the centre of the dislocation from the position shown in figure 1, the energy becomes

$$\frac{1}{2}\frac{\mu d^2}{4\pi^2}\sum_{n=-\infty}^{\infty}[\text{const.} + \cos 2\{\tan^{-1} 2(1-\sigma)(\alpha + \tfrac{1}{2}n)\}]. \qquad \ldots\ldots(31)$$

This may be transformed by using the relation

$$\sum_{-\infty}^{\infty} f(n) = \sum_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) \cos 2\pi xs\, dx \qquad \ldots\ldots(32)$$

into

$$\frac{\mu d^2}{8\pi^2}\int_{-\infty}^{\infty} [\text{const.} + \cos 2\{\tan^{-1} 2(1-\sigma)(\alpha + \tfrac{1}{2}x)\}]\, dx$$

$$+ \frac{\mu d^2}{4\pi^2}\sum_{s=1}^{\infty}\int_{-\infty}^{\infty} [\text{const.} + \cos 2\{\tan^{-1} 2(1-\sigma)(\alpha + \tfrac{1}{2}x)\}] \cos 2\pi xs\, dx.$$

Writing

$$z = 2(1-\sigma)(\alpha + \tfrac{1}{2}x)$$

reduces this to

$$\frac{\mu d^2}{8\pi^2(1-\sigma)}\int_{-\infty}^{\infty} [\text{const.} + \cos\{2\tan^{-1} z\}]\, dz$$

$$+ \frac{\mu d^2}{4\pi^2(1-\sigma)}\sum_{s=1}^{\infty}\int_{-\infty}^{\infty} [\text{const.} + \cos\{2\tan^{-1} z\}] \cos 2\pi s\left(\frac{z}{1-\sigma} - 2\alpha\right) dz.$$

The first term is an infinite constant independent of α, and the second reduces to

$$\frac{\mu d^2}{2\pi^2(1-\sigma)}\sum_{s=1}^{\infty} \cos 4\pi s\alpha \int_{-\infty}^{\infty} \cos\left(\frac{2\pi s z}{1-\sigma}\right)\frac{dz}{1+z^2}. \qquad \ldots\ldots(33)$$

The integral in (33) is equal to $\int_{-\infty}^{\infty} e^{ikz}dz/(1+z^2)$, which is readily shown by contour integration to be πe^{-k}, where $k = 2\pi s/(1-\sigma)$. The energy is then

$$v = \frac{\mu d^2}{2\pi(1-\sigma)}\sum_{s=1}^{\infty} e^{-\frac{2\pi s}{1-\sigma}} \cos 4\pi s\alpha. \qquad \ldots\ldots(34)$$

The work done in an infinitesimal displacement δ of the dislocation is

$$\frac{\delta}{d}\frac{dv}{d\alpha} = -\frac{2\mu d\delta}{1-\sigma}\sum_{s=1}^{\infty} s e^{-\frac{2\pi s}{1-\sigma}} \sin 4\pi s\alpha.$$

The term in $s=1$ dominates this sum, and has a maximum value

$$\frac{2\mu d\delta}{1-\sigma} e^{-\frac{2\pi}{1-\sigma}}. \qquad \ldots\ldots(35)$$

The work done by an external shearing stress T in this displacement is $Td\delta$, and by comparison with (35) it follows that the smallest external stress which will cause a single dislocation to move continuously through the lattice is given by

$$T = \frac{2\mu}{1-\sigma} e^{-\frac{2\pi}{1-\sigma}}, \qquad \ldots\ldots(36)$$

whereas the shearing stress at which one lattice plane moves rigidly over another is $\mu/2\pi$. The ratio of the stress required to move a dislocation to the theoretical shear strength of a perfect lattice is

$$\frac{4\pi}{1-\sigma} e^{-\frac{2\pi}{1-\sigma}}. \qquad \ldots\ldots(37)$$

Since the width of the dislocation given by (12) is only half that given by (3) in the original note, the energy of a dislocation is more sensitive to its position in the lattice, and the ratio (37) is of a larger order of magnitude than the original estimate. Values are:

Poisson's ratio	0·2	0·3	0·4
Ratio of stresses	6×10^{-3}	2×10^{-3}	6×10^{-4}

Expression (37) differs from Peierls's expression not only in replacing $(1-\sigma)$ by $2(1-\sigma)$, but also in the form of the coefficient. This is because Peierls replaces the integrals (19) and (25) by the corresponding sums, whereas in (37) only (25) has been replaced by a sum. The order of magnitude of the result is the same on each approximation, the exponential factors agreeing. Any calculation of the effect of atomic structure which is based on the result (11) of the theory of a continuous medium is really inconsistent, and the answer is therefore undefined.

The shear stress T given by (36) would cause a uniform shear of the lattice through a small angle $\theta = T/\mu$, and this stress would just equal the mutual attraction of a pair of dislocations of unlike sign in the same glide plane at a distance apart of $2d/(1-\sigma)\theta$. To this approximation, a pair of dislocations in an unstressed crystal will coalesce if they are close together, but their attraction will not overcome the forces anchoring them to their places in the lattice if their separation exceeds

$$de^{\frac{2\pi}{1-\sigma}}. \qquad \text{......(38)}$$

Numerically this implies:—

Poisson's ratio	0·2	0·3	0·4
Critical separation	$2500\,d$	$8000\,d$	$35000\,d$

§7. DISCUSSION

To the approximation considered here, dislocations in a lattice have the following properties. The width of a dislocation is small, displacements comparable with the interatomic distance being confined to a few atoms. The energy required to form a pair of dislocations in a crystal under moderate shear stress is large, of the order of 10 electron volts for each atomic plane. The energy required to form even a short dislocation far exceeds the thermal energy of an atom at room temperature. A single dislocation will not move freely through a lattice under very small stresses, but will do so under shear stresses of the order of a thousandth of the theoretical shear strength of the perfect lattice. This figure is of the same order of magnitude as that commonly accepted as the elastic limit of a real single crystal. Two dislocations of opposite sign in the same glide plane attract each other with a force which at large distances is inversely proportional to the distance between them, and they will run together and coalesce unless they are separated by a distance of the order of 10000 lattice spacings. This spacing is of the same order as the size of the mosaic blocks of which real crystals are usually composed, and may represent the reason for the appearance of this characteristic length in the growth of real crystals.

It is, however, necessary to consider how far this approximate treatment of a simplified model can be expected to represent the properties of a real crystal. The first approximation is the use of a simple cubic lattice. Glide processes are best understood in face-centred cubic and hexagonal metals. In both cases glide takes place in a close-packed direction over close-packed planes, but the actual geometrical form of the dislocation in such lattices is not known or easy to visualize. It seems unlikely that this approximation should greatly affect the width or the energy of a dislocation, but the stress required to move a dislocation, and the critical separation of two dislocations, which depend very sensitively on the width of a dislocation, might be entirely different.

The next approximation is the assumption that the crystal is isotropic. It is easy to show (cf. Appendix) that the elastic modulus for shearing a cubic crystal in the direction $(1\bar{1}0)$ across a (111) plane is $4[s_{44}+\frac{1}{3}(s_{11}-s_{12}-2s_{44})]$. The anisotropy term $\frac{1}{3}(s_{11}-s_{12}-2s_{44})$ is only 20% of s_{44} for Cu, Ag and Au, so it seems likely that anisotropy will not be of great importance. Here again the stress required to move a dislocation depends critically on its width.

Probably the most serious assumption is the sinusoidal law of force (1). In the majority of metals, the ionic shells are in contact, and for many purposes it is satisfactory to regard their lattices as composed of hard spheres in contact. Dr. Orowan has pointed out in discussion that such an extreme model does not have an elastic modulus in the usual sense, because a finite force is required to produce an infinitesimal displacement of two neighbouring layers of atoms. If this crude picture is refined somewhat by considering the spheres to be held apart by forces which vary very much more rapidly with distance than the attractive forces, the work required to slide one layer of atoms a distance $\frac{1}{2}d$ over the next layer is still much less than μd^2, which is the order of magnitude corresponding to (1). It follows from the discussion of § 1 that if in fact the surface energy is much less than that corresponding to a sine law, the width of a dislocation will be greater than that calculated, since an increase in the width reduces its elastic energy without a corresponding increase in its surface energy. Its total energy is less (but could hardly be so much reduced that pairs of dislocations could be formed by thermal agitation) and it moves more easily.

The following argument, due to Professor Mott, shows that this correction is not large. Consider a two-dimensional model, which in its equilibrium state consists of close-packed cylinders (figure 4*a*). In the dislocated state the cylinders are arranged as in figure 4*b*.

Let the distance between the layers of atoms on either side of the slip plane be z, while each atom in the upper row is displaced horizontally a distance ϕ from its neighbour in the lower row. It is assumed in the argument that the only forces acting are a long-range force, the potential of which depends only on the volume of the crystal, and short-range repulsions between neighbouring ions. This is not correct, for the alkali metals, in which the short-range repulsion is small, have rigidities comparable with their bulk moduli. These neglected forces are such that the configuration of figure 4*a* is one of stable equilibrium; when the magnitude of the other forces is estimated from the observed rigidity of the crystal under small displacements it is therefore over-estimated. The neglected forces depend on long-range interactions alone, and are consequently smoothly varying functions of ϕ which cannot depart far from a sinusoidal form. The estimate of the departure of the law of force from a sinusoidal form which is made by neglecting these forces must exceed the true departure.

It will be convenient to represent the energy of the cohesive forces by an expression of the form Bz^2, that is to say, proportional to V^2, where V is the atomic volume. In fact the electrostatic energy is proportional to $V^{-\frac{1}{3}}$, and is opposed by the Fermi energy, which is proportional to $V^{-\frac{2}{3}}$. The constant B is determined by the condition that the long-range force should equal the short-range repulsion at the equilibrium interatomic distance. It may be shown that for larger values of V the deviation of the true energy from that given by the approximate expression Bz^2 is greatest when the Fermi energy is neglected, and that if the Fermi energy is neglected the approximate formula over-estimates the difference of the energies of figures 4 *a* and 4 *b* by less than 20%. The repulsive potential, which falls off rapidly with distance, is taken to be of the form $Ae^{-\beta r^2}$, where $\beta d^2 \gg 1$.

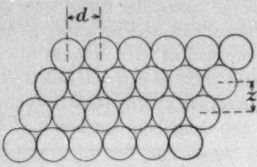


Figure 4 *a*. A normal crystal ($\phi = \frac{1}{2}d$).

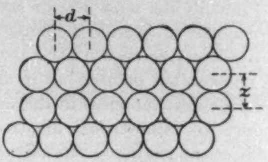


Figure 4 *b*. A dislocated crystal ($\phi = 0$).

The energy per atom is then

$$W = Bz^2 + A \sum_{n=-\infty}^{\infty} e^{-\beta[(nd+\phi)^2+z^2]}. \qquad \ldots\ldots(39)$$

For each value of ϕ, the equilibrium value of z is given by $\partial W/\partial z = 0$, which leads to

$$W = \frac{B}{\beta}\left[1 + \log\left\{\frac{\beta A}{B}\sum_n e^{-\beta(nd+\phi)^2}\right\}\right]. \qquad \ldots\ldots(40)$$

For figure 4 *a*, $\phi = \frac{1}{2}d$, and, neglecting the repulsions between all except nearest neighbours,

$$W_a = \frac{B}{\beta}\left[1 + \log\frac{2\beta A}{B} - \tfrac{1}{4}\beta d^2\right].$$

Similarly $$W_b = \frac{B}{\beta}\left[1 + \log\frac{\beta A}{B}\right].$$

The energy difference is

$$W_b - W_a = \frac{B}{4\beta}[\beta d^2 - 4\log 2]. \qquad \ldots\ldots(41)$$

Near the position of stable equilibrium (40) may be written approximately

$$W = \frac{B}{\beta}\left[1 + \log\frac{\beta A}{B} + \log\left\{e^{-\beta(d-\phi)^2} + e^{-\beta\phi^2}\right\}\right].$$

This gives

$$\frac{dW}{d\phi}=2B\frac{(d-\phi)e^{-\beta(d-\phi)^2}-\phi e^{-\beta\phi^2}}{e^{-\beta(d-\phi)^2}+e^{-\beta\phi^2}},$$

which vanishes for $\phi=\tfrac{1}{2}d$, and when $\phi=\tfrac{1}{2}d$,

$$\frac{d^2W}{d\phi^2}=B(\beta d^2-2). \qquad \ldots\ldots(42)$$

For a sinusoidal law of force of the form

$$W=P+Q\cos 2\pi\phi/d$$

the relations corresponding to (41) and (42) are

$$W_b-W_a=2Q \qquad \ldots\ldots(43)$$

$$\frac{d^2W}{d\phi^2}=\frac{4\pi^2Q}{d^2}. \qquad \ldots\ldots(44)$$

If now Q is determined by a comparison of (42) and (44), which is the method used in §2, its value is found to be

$$Q=Bd^2(\beta d^2-2)/4\pi^2,$$

and the energy of each atom in the dislocated plane exceeds that of a normal atom by the amount given by (43), namely,

$$W_b-W_a=Bd^2(\beta d^2-2)/2\pi^2. \qquad \ldots\ldots(45)$$

This value (45) exceeds the true value (41) by a factor

$$\frac{2\beta d^2}{\pi^2}\cdot\frac{\beta d^2-2}{\beta d^2-4\log 2}. \qquad \ldots\ldots(46)$$

For very hard ions $\beta d^2 \gg 1$ and the factor (46) is large. The value of (46) for copper may be estimated from the values of dW/dr and d^2W/dr^2 given by Fuchs (1935). Here W is the repulsive energy, which in this calculation is assumed to be given by $W=Ae^{-\beta r^2}$. It follows that

$$\beta r^2=\tfrac{1}{2}-r\frac{d^2W}{dr^2}\bigg/2\frac{dW}{dr},$$

and Fuch's numerical values lead to $\beta r^2=20{\cdot}3$ at the equilibrium distance $r=d$. The value of the factor (46) is then 4·25. The sinusoidal law (1) therefore overestimates the energy of a surface of misfit by a factor of less than 4·25. The corresponding errors in estimating the size and energy of a dislocation are at most factors of about 2.

It seems unlikely that the purely mathematical approximations should greatly influence the result. The most dangerous assumption seems to be that of §6, which is that the displacements of the individual atoms are given exactly by (12) for all positions of the dislocation in the lattice. A more detailed investigation might show that the second approximations for the displacements of the atoms in the extreme positions of the dislocation in the lattice (corresponding to $\alpha=0$ and $\alpha=\tfrac{1}{4}$ in (31)) lead to differences in the energies of these two positions which cannot be neglected in comparison with the very small difference calculated by using (12) in each case. It is difficult to see how to obtain such a second approximation, or to decide on general grounds whether the reduction in energy corresponding to the second approximation would be greater for the stable position ($\alpha = 0$) or the unstable position ($\alpha = \frac{1}{2}$).

ACKNOWLEDGMENTS

The writer's thanks are due to Professor Peierls for an explanation of the methods used in deriving the results previously published, and for permission to quote sections of his paper. He is also indebted to Professor Mott for valuable discussions, and to Dr. E. H. Linfoot for pointing out some of the questions of convergence involved in the derivation of equation (10).

APPENDIX

THE CONTRIBUTION OF ANISOTROPY TO THE SHEAR MODULUS IN THE SLIP DIRECTION

Consider the stress system which, referred to the principal axes of a cubic crystal, is represented by

$$\begin{bmatrix} 2 & 0 & 1 \\ 0 & -2 & -1 \\ 1 & -1 & 0 \end{bmatrix} \times \frac{S}{6^{\frac{1}{2}}}. \qquad \ldots\ldots(\mathrm{A}1)$$

This is easily shown to represent a shearing stress S across the plane (111) in the direction $(1\bar{1}0)$.

The corresponding strain in a material of cubic symmetry is

$$\begin{bmatrix} 2(s_{11}-s_{12}) & 0 & 2s_{44} \\ 0 & -2(s_{11}-s_{12}) & -2s_{44} \\ 2s_{44} & -2s_{44} & 0 \end{bmatrix} \times \frac{S}{6^{\frac{1}{2}}}. \qquad \ldots\ldots(\mathrm{A}2)$$

The component of shear strain in the direction corresponding to the stress is

$$2Ss_{44} + \tfrac{2}{3}S(s_{11} - s_{12} - 2s_{44}),$$

and the angle of shear is twice this.

REFERENCES

Burgers, J. M., 1940. *Proc. Phys. Soc.*, **52,** 23.
Dehlinger, V., 1929. *Ann. Phys., Lpz.*, **2,** 749.
Fuchs, K., 1935. *Proc. Roy. Soc.*, A, **153,** 622.
Koehler, J. S., 1941. *Phys. Rev.*, **60,** 397 (equation 18).
Love, A. E. H., 1944. *Elasticity* (Cambridge : The University Press) (§ 144).
Peierls, R., 1940. *Proc. Phys. Soc.*, **52,** 34.
Taylor, G. I., 1934. *Proc. Roy. Soc.*, A, **145,** 362.
Titchmarsh, E.C., 1937. *Fourier Integrals* (Oxford : The Clarendon Press), Chap. V.

A NOTE ON THE EFFECT AT THE CATHODE OF AN ARC BETWEEN COPPER ELECTRODES

By MAURICE MILBOURN,
Imperial Chemical Industries Ltd., Metals Division

MS. received 19 *September* 1946

ABSTRACT. Observations on burning arcs and on arced electrodes have indicated that melting of a copper cathode does not necessarily take place, and that selective distillation of impurities occurs when melting is induced by the presence of a metal having powerful reducing properties. Volatilization of copper from the cathode appears to be effected through the formation of cuprous oxide.

§1. INTRODUCTION

In a normal copper arc, as used for spectrographic analysis, volatilization takes place at a greater rate from the negative than from the positive electrode. Selective distillation of impurities from the cathode does not occur to any great extent, but elements with low boiling points may be removed preferentially if free oxidation of the negative electrode is hindered by the presence of a metal having powerful reducing properties. At the same time, the rate of volatilization of copper is reduced (Milbourn, 1943).

It has been suggested (Milbourn, 1944) that impurities distil out of the electrode material only when an appreciable volume of metal is molten, and that the absence of selective distillation indicates that no such melting is taking place. This point of view cannot easily be reconciled, however, with generally accepted theories of the arc discharge, which postulate that the cathode is at or near the boiling point of the metal. For instance, it has been estimated by von Engel and Steenbeck (1934) that a copper cathode attains a temperature of about 2000° K. At the same time, layers of oxide seem to be necessary for the maintenance of a normal arc, so that a reasonable energy balance may be established (von Engel, 1935).

§2. EXPERIMENTAL OBSERVATIONS

In order to obtain information on the volatilization of material, the degree of melting, and the distribution of oxide layers in copper cathodes, observations have been made on arcs while they were burning, and on electrodes after they had been subjected to an arc 2 to 3 mm. in length and carrying a current of 5 amp. for 30 to 60 seconds, conditions which could very well be used for analytical purposes.

Examination of a burning copper arc with a low-power binocular microscope, using a suitable filter, showed that the cathode spot is more or less surrounded by a rim of molten material, but that it leaves behind it, as it travels, an area which is free from such material. It appears, in fact, that the cathode spot travels towards regions where molten material has collected and vaporizes it.

Sections cut through the small area on the negative electrode where the arc was striking when it was extinguished have been examined microscopically, some of the results being illustrated in the accompanying photographs. Figure 1 shows a section of a ½ in. diameter pure copper rod, on which the arc passed to a pyramidal point. The shallow depression in the surface of the specimen is a section of the area where the cathode spot was acting when the arc was extinguished, and although there is no apparent layer superimposed on the copper in the depression, there is a rim of bluish-coloured cuprous oxide on each side of it. These rims form the molten material observed on the surface of the electrode while the arc was running. The fact that the crystal structure of the copper is undisturbed up to the point of striking of the arc indicates that very little, if any, melting of the metal itself has taken place, and that its temperature was probably considerably below 1000° c. Figure 2 is taken from a similar section through an arced sample of copper strip, about 0·020 in. thick. The oxide rim can again be seen, but in this case the smaller dimensions of the sample have allowed the material to become hotter, although melting has not occurred to any detectable extent.

A strip of 93/7 copper-aluminium alloy, illustrated in figure 3, shows very considerable melting and little oxidation under similar conditions. It is clear that selective distillation would be expected from this sample to a much greater degree than from the copper samples illustrated in figures 1 and 2. The observations on melting thus confirm those made previously on the relative amounts of selective distillation from copper and from small samples containing a powerful reducing agent. The presence of aluminium has prevented oxidation of the copper, so that the heat of vaporization of copper oxide, as well as the heat of formation of aluminium oxide, are made available for melting the basis metal.

By moving a flat copper surface horizontally at about one inch per second, while an arc was striking it vertically, a continuous streak was obtained showing the instantaneous state of the cathode, illustrated in figure 4.

The central region of the arc streak has well-defined, although irregular, edges, and the area between them is covered with a very thin layer of cuprous oxide, which is readily soluble in dilute hydrochloric acid, together with a few strings of globules of the same material. Both of these are firmly adherent, but the material outside this central region can be easily removed with dry cotton wool. The evidence is that the cuprous oxide layer is formed while the arc is passing, and that it collects firstly into small globules and then into comparatively thick layers at the edge of the arc, from whence it is volatilized. It may be noted that cupric oxide decomposes into cuprous oxide at 1200° c., and that the latter melts at 1235° c., eventually decomposing into copper and oxygen at 1800° c. The equilibrium of various compounds at high temperature may therefore play a very important part in the functioning of an arc.

§ 3. CONCLUSIONS

These observations show that melting of the negative electrode does not necessarily take place during the running of a normal arc between copper electrodes and, if it does, it can be an extremely localized effect. Selective distillation is liable to occur when the basis metal is molten, but volatilization of copper is

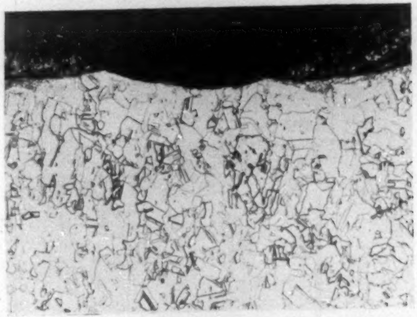


Figure 1. Section through cathode spot on copper rod. × 100.

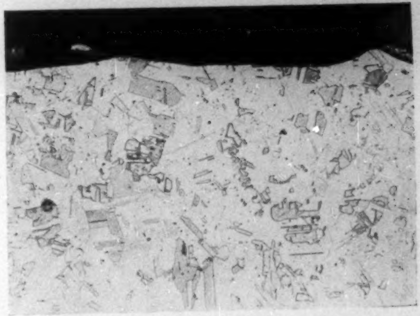


Figure 2. Section through cathode spot on copper strip. × 100.

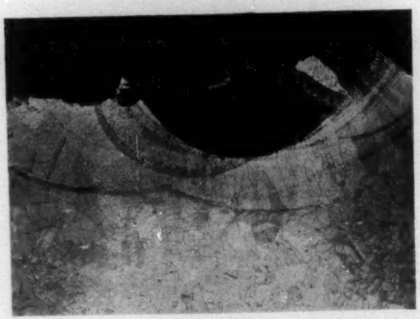


Figure 3. Section through cathode spot on aluminium bronze. × 20.

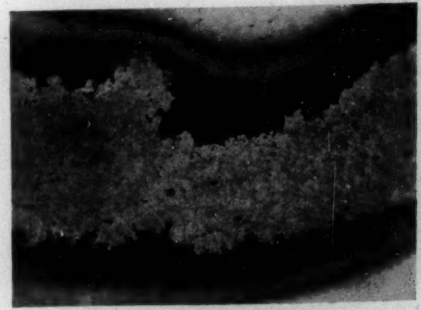


Figure 4. Arc streak on copper cathode. × 65.

effected through the formation of cuprous oxide. These observations have a direct bearing on the spectrographic analysis of copper and its alloys by an arc technique, but they may also be of value in connection with other types of spectrographic source, which include arc-like phases.

§4 ACKNOWLEDGMENT

The author's thanks are due to Dr. Maurice Cook for his helpful advice and interest in the work.

REFERENCES

von Engel, A. and Steenbeck, M., 1934. *Elektrische Gasentladungen*, vol. ii, p. 136.
von Engel, A., 1935. *Wiss. Veröff. SiemensWerken*, **14**, 38.
Milbourn, M., 1943. *J. Inst. Metals*, **69**, 441–463.
Milbourn, M., 1944. *J. Inst. Metals*, **70**, 299.

DAMPING CAPACITY, STRAIN HARDENING AND FATIGUE

By R. F. HANSTOCK,
High Duty Alloys Ltd.

MS. received 17 *August* 1946

ABSTRACT. An electromagnetic method of exciting torsional resonance vibrations is described.

For some alloys of aluminium, notably binary alloys containing 5% and 11% of magnesium, vibrational strains of sufficient magnitude to cause fatigue cracks can be developed by this method at frequencies of the order of one kilocycle per second. During vibration, measurements are made of the damping capacity of the specimens to provide information concerning strain hardening and energy conversion within the specimens up to the time when failure occurs by fatigue.

Fatigue failure of the two binary alloys containing magnesium is shown to be preceded by strain hardening. The endurance of individual specimens is related to the amount of irreversible strain imposed per cycle, failure occurring when the cumulative internal strain approaches a limiting value.

§1. INTRODUCTION

When a polycrystalline metal is subjected to alternating stresses, the material may fail by fracture after a certain number of stress cycles even though the maximum stress involved is considerably lower than the static stress required to cause rupture. Described in general terms as fatigue, this phenomenon is investigated conventionally by subjecting a specimen of the metal to an alternating stress (frequency normally about 100 cycles per second), the number of stress cycles required to cause fracture being recorded. A series of tests of this type, over a range of maximum values of the alternating stress, gives a fatigue curve for the material.

Strain hardening is a result of plastic or irreversible straining of the metal and is characterized by an increase in the stress necessary to produce a defined amount of plastic strain, i.e. as an increase in the limit of proportionality or proof stress. During strain hardening a part of the work done is stored in the metal and subsequently may be released by annealing.

Damping capacity, or internal friction, refers to the capacity of a solid to convert vibrational energy to some other form. It is defined normally as $\Delta E/E$, where ΔE is the energy absorbed and/or dissipated during each cycle of the vibration, and E is the total mechanical energy stored within the material for defined conditions of vibration.

Fatigue and strain hardening are the results of energy conversion, whilst damping capacity expresses the ability of the material to cause this conversion. The object of this paper is to show the relation of damping capacity to strain hardening and fatigue by referring to an experimental examination of two alloys of aluminium and magnesium.

This work became possible as a result of the development of a method of measuring damping capacity during vibration tests on specimens at strains within the fatigue range. The method is described in an earlier paper (Hanstock and Murray, 1946) which also records that damping capacity is a function of the maximum vibrational surface strain, ϕ, of the metallurgical state and of the vibrational history of the specimen. It is also known, qualitatively, that $\Delta E/E$ is sensitively dependent on the temperature of the specimen when a certain temperature, probably characteristic of the material, is exceeded. In the present investigation, the temperature of the specimens was kept within the region where $\Delta E/E$ is insensitive to variations of temperature.

§2. FORM OF SPECIMEN AND METHOD OF EXCITATION

The specimens have the form of the solid of revolution, about the vibration axis, XY, of the section shown in figure 1. The mode of vibration about XY is torsional, PQ being the nodal plane. Sections AB and CD are "inertias" and section BC is the elastic member of the system. The natural frequency of such a specimen (of alloys consisting mainly of aluminium) is about 960 cycles per second, and this is the frequency at which the fatigue tests are made.*

The specimen, S, figure 2, is suspended freely by a fine steel wire, W (0·018 cm. diameter and approximately 20 cm. long), the vibrational axis being vertical. At the junction of the wire and the specimen there is a small mirror, M, which reflects a beam of light from the optical system, L, on to a graduated scale (not shown). From the spread of the beam of light on the scale, the amplitude of vibration and the maximum surface strain on the elastic section of the specimen may be calculated. The mirror is brought into a suitable position for reflection by rotating the wheel, R. Vibrations at resonance frequency are excited by applying a simple harmonic couple electromagnetically to the upper end of the specimen. No mechanical coupling is involved, so that conversion of vibrational energy, other than as a result of the internal friction of the specimen, is very small.

---

* The actual frequency is measured during the test by means of a calibrated beat-frequency oscillator.

The following is a brief description of the method of excitation, but reference should be made to an earlier paper (*loc. cit.*) for a more detailed description of the electrical equipment. Two coils, C, figure 2, surround but do not touch the lower inertial section of the specimen. A magnetic field, originating from the permanent magnet, P, is so disposed in relation to these coils that, when the inertia is in alternating angular motion about the vibration axis, an induced current flows within this part of the specimen and causes an alternating e.m.f. to be developed in the coils. This e.m.f., which is of the same frequency as, and is proportional to the amplitude of the motion of, the specimen, forms the signal input of an amplifying system. The output of the amplifier (50 watts maximum) passes to coils, E, surrounding but not touching the upper inertial section of the specimen. The current induced in this inertia reacts with a magnetic field (produced by energizing the coil, F) to generate an alternating couple acting on the specimen.

Thus the whole system comprises an electromechanical oscillator, the frequency of which is determined by the dimensions and nature of the specimen, whilst the amplitude of the mechanical vibration depends on the damping capacity of the specimen and on the power supplied by the amplifier. The lower coil system, or detector, is magnetically shielded; part of the shielding has been removed in figure 2 to allow the coils to be seen. The upper coil system, or exciter, is air cooled through pipes, A, to prevent the temperature of the specimen rising to within the region where $\Delta E/E$ becomes dependent on temperature.

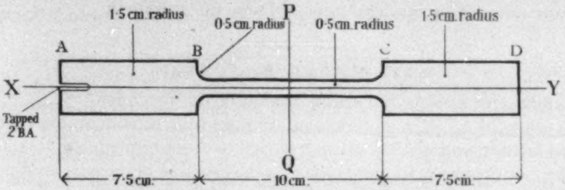


Figure 1. Section of specimen in plane containing vibration axis X, Y.

§3. MEASUREMENT OF DAMPING CAPACITY

For suitably constructed exciter and detector coils it may be shown that $\Delta E/E$ is directly proportional to I/V (to within 3% for aluminium alloys), where I is the R.M.S. current in the exciter and V is the R.M.S. voltage developed in the detector. For aluminium alloys it has been shown (*loc. cit.*) that $\Delta E/E$ is substantially independent of variation of the maximum surface shear strain for strains generally up to 1×10^{-4} and very often, depending on the composition and metallurgical state of the alloy, up to strains considerably in excess of this value. Consequently it is permissible to determine $\Delta E/E$ within this range of strain by measuring the logarithmic decrement, λ, of free decay of the vibration. Provided that $\Delta E/E$ is small, $\Delta E/E = 2\lambda$. One absolute determination of $\Delta E/E$ from a measurement of the half-amplitude decay time at low strains is sufficient to allow the proportionality

$$\Delta E/E \propto I/V$$

to be used for the determination of $\Delta E/E$ from I/V at higher surface strains during steady vibration of the specimen.

The half-amplitude decay time at low strains is determined by means of a stop-watch, the decay of amplitude being observed as a decay of the detector voltage. For specimens of the type under consideration, the half-amplitude decay time is about 29 seconds, which corresponds to a damping capacity of 5×10^{-5}. In order to simplify observations during fatigue tests, the ratio I/V is determined directly by potentiometric comparison of the detector voltage and of the voltage drop across a one-ohm resistance carrying the exciter current.

Having determined $\Delta E/E$, the following quantities may be calculated:—
Energy dissipated per cycle

$$\Delta E = \Delta E/E \, . \, \tfrac{1}{4} nv\phi^2 \text{ ergs.} \qquad \ldots\ldots(1)$$

where n = modulus of rigidity (dyne/cm.^{2});
v = volume of the elastic section of the specimen;
ϕ = peak value of the alternating surface strain on the elastic section of the specimen (radians).

Maximum value of the alternating couple on the specimen

$$G = \frac{\Delta E}{E} \, . \, \frac{nr^3\phi}{2} \text{ dyne cm.} \qquad \ldots\ldots(2)$$

where r = radius of the elastic section of the specimen (cm.).

The values of $\Delta E/E$, as determined, represent the total damping capacity per cycle and include energy dissipations external to the specimen (e.g. losses through and in the specimen wire, frictional losses at the junction of the wire and specimens, air friction losses and losses due to electromagnetic damping).

The sum of these losses can be estimated experimentally, and in the present work is found to be represented by a damping capacity of 2×10^{-5}, which is independent of the amplitude of the vibration.

§4. THE ALLOYS INVESTIGATED

It will be appreciated that with an electrical amplifying system of limited output, the maximum vibrational strain which can be developed in the mechanical vibrator will depend on the damping capacity of the material. During a series of experiments with various alloys of aluminium, it was found that binary alloys containing 5% and 11% of magnesium had a low and practically constant damping capacity for surface shear strains up to about 15×10^{-4} and 30×10^{-4} respectively, above which values the damping capacity began to increase. This relation between damping capacity and strain is typical of aluminium alloys, but the range of strains over which the damping capacity remains small and constant is unusually extensive for the alloy containing 11% magnesium. Further, these alloys are known to be capable of appreciable strain-hardening and, as the results to be given later will show, can be strain-hardened by vibration, thus causing the damping capacity at any strain, greater than the respective values given above, to decrease during vibration. It appeared, therefore, that it would be possible to excite vibrations of sufficient amplitude to produce fatigue cracks in these alloys, even though the maximum power available did not exceed 50 watts.

Normally, the metallurgical state of such alloys would be made definite by raising them to a temperature of 430° C. for sufficient time to ensure complete solution of the Al_2Mg_3 compound and by rapid cooling to retain this compound in solid solution; alternatively they might be cooled slowly from 430° C. to allow the precipitation of Al_2Mg_3. The separation of Al_2Mg_3 causes a considerable reduction in ductility. In the present investigation the metallurgical state of the material was not of the first importance. Primarily it was necessary to obtain a number of specimens all in a similar stable condition. The specimens were taken from $3\frac{1}{2}$ in. diameter extruded bar, stored for two years after extrusion so that no marked differences in the metallurgical states of the individual specimens would be expected. Metallographic examination of the alloys showed that the Al_2Mg_3 was in solution in the 5% alloy but not entirely so in the 11% alloy. A few specimens of the 11% alloy, heat-treated to obtain complete solution, were found to behave not very differently from the alloy as extruded, when subjected to the vibration tests.

The specimens were prepared by careful machining to the dimensions shown in figure 1, from pieces sawn off-centre from the $3\frac{1}{2}$ in. diameter bar. After machining, the central section of each specimen was polished.

§ 5. METHOD OF EXAMINATION

In conventional fatigue tests, the specimen is subjected to stress or strain alternations between certain defined limits, the cyclic process being continued until a fatigue fracture occurs. It is not convenient to adopt these conditions for the operation of the electromechanical system owing to the limited output of the amplifier.

In many cases the damping capacity of the specimen changes during vibration. If it increases, the output of the amplifier must be increased to maintain a constant amplitude of vibration and eventually the maximum output may be insufficient to maintain the desired amplitude. If the amplitude decreases during vibration, as is the case for the alloys now being considered, the output must be decreased to maintain a constant amplitude, and consequently the maximum available power will not be employed except for short periods at the commencement of the test; under such conditions the equipment will not be used efficiently for production of fatigue.

In practice, the amplifier is operated to maintain a constant value of the R.M.S. current output whilst testing any one specimen. This operating condition is such that the peak value of the alternating couple applied to the specimen remains constant throughout the test.

The course of a typical test is as follows:—The specimen is vibrated at an amplitude corresponding to a maximum surface strain of $0{\cdot}75 \times 10^{-4}$ and the I/V ratio is measured. The strain is then increased to $1{\cdot}5 \times 10^{-4}$, excitation being stopped when this value is reached. As the amplitude decays, the time is measured for the strain to fall from $1{\cdot}0 \times 10^{-4}$ to $0{\cdot}5 \times 10^{-4}$. From this decay time, the damping capacity corresponding to an average strain of $0{\cdot}75 \times 10^{-4}$ is calculated. The specimen is then vibrated at the selected value of the exciter current and observations are made of the amplitude and of the I/V ratio as the test proceeds, until either the specimen fails by fatigue or the number of vibration cycles exceeds 10^8. The frequency of the specimen is determined periodically during the test.

§6. RESULTS

Alloy containing 11% *magnesium*

The experimental observations on the 11% alloy are summarized in figures 3 and 4, which show respectively, the change in surface shear strain (i.e. amplitude) and the change in damping capacity during vibration up to fracture, or to more than 10^8 cycles, of several specimens operated on by couples of peak values ranging from 4 gm./cm. to 100 gm./cm.

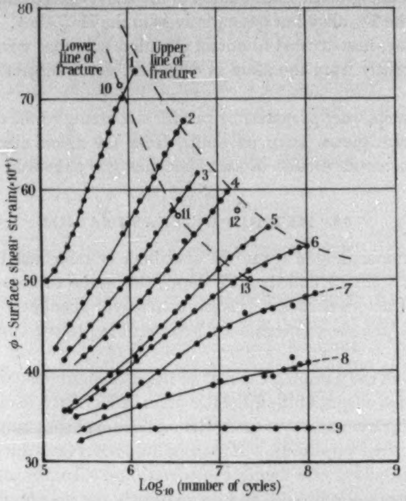


Figure 3. **Variation of strain (amplitude) during vibration of an alloy of aluminium containing 11% Mg.**

Increase in strain and decrease in damping capacity occur during vibration, the rate of change depending on the magnitude of the alternating couple. The formation of a fatigue crack is indicated sharply by a decrease in amplitude, an increase in damping capacity, and a fall in frequency of up to 100 cycles per second, depending on the size and disposition of the crack. Figure 5 shows a typical fracture exhibiting the usual characteristics of failure by fatigue. The specimens do not fracture completely during vibration, but can be broken easily afterwards, since the crack normally extends at least to the vibration axis. The rate at which the crack is formed is fairly high, a crack of the dimensions shown in figure 5 occurring in about 2 minutes (10^5 cycles).

The upper broken curve in figure 3 is a type of fatigue curve but differs from the conventional type in that the conditions it represents are not constant stress or

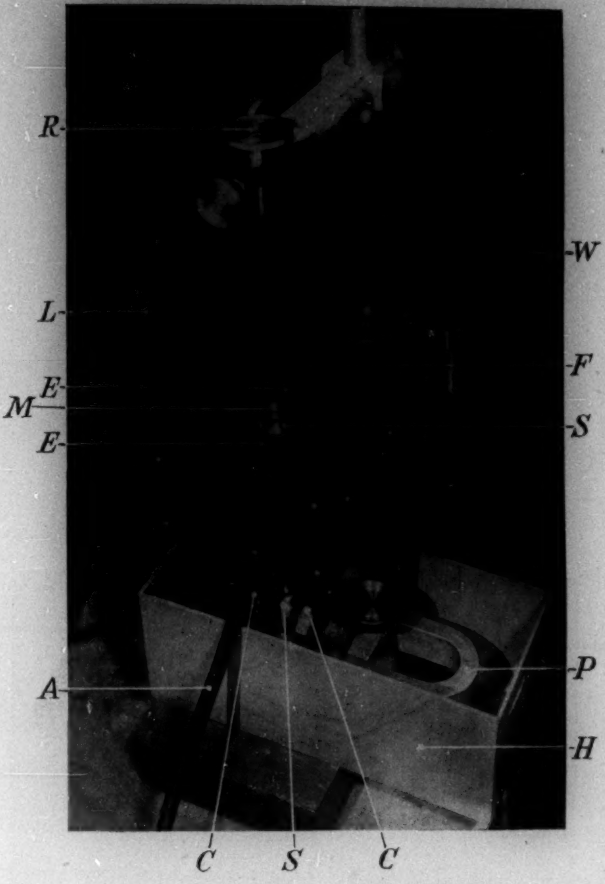

R
W
L
F
E
M
E
S
A
P
H
C
S
C

Figure 2. Exciter and detector units for torsional vibrations.

To face page 280

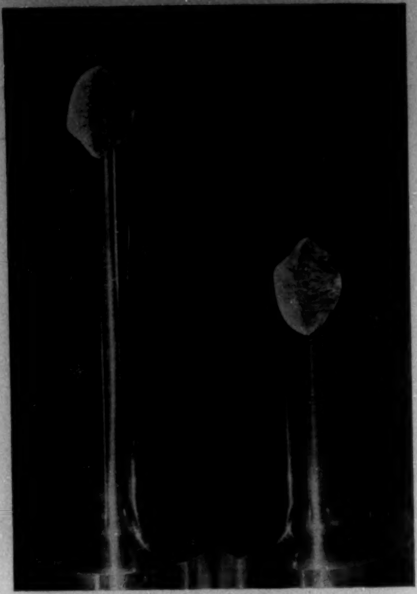


Figure 5. Fatigue fracture of aluminium alloy containing 11% Mg.

3% Mg
5% Mg
11% Mg
4% Cu
4% Cu

Figure 8. Fatigue cracks in aluminium alloys.

strain for each test but constant value of the alternating couple applied to a resonant system.

Further observed and derived data are given in table 1.

Table 1

Specimen number	Frequency (c.p.s.)	Couple (gm. cm.)	Total no. of cycles	$\Sigma_0^N \Delta E$ (ergs)	Mean ΔE (ergs)
1	930	100	$1{\cdot}23 \times 10^6$	$2{\cdot}29 \times 10^{10}$	18650
2	954	40	4·3	3·10	7200
3	965	25	6·0	2·30	3830
4	959	15·5	13·0	3·00	2310
5	973	11·2	37·1	5·15	1390
6	969	8·7	101	10·9	1080
7	975	7·4	124*	10·2	825
8	928	6·0	103*	5·74	556
9	962	4·0	103*	3·07	298

* Specimen not fractured.

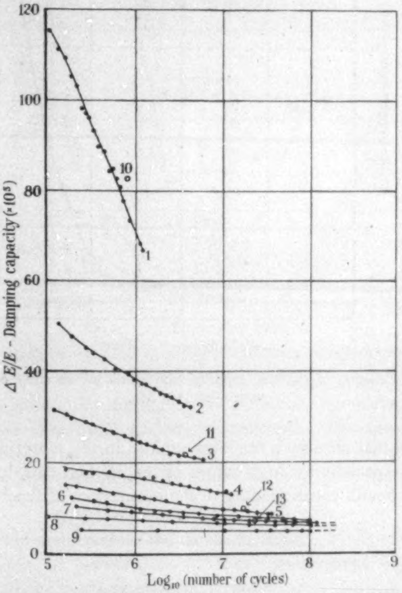


Figure 4. Variation of damping capacity during vibration of an alloy of aluminium containing 11% Mg.

From figures 3 and 4, graphs may be drawn showing the variation of ΔE with N, and from these graphs the total energy dissipated at the end of the test may be estimated. This energy is shown as $\Sigma_0^N \Delta E$ in table 1; the mean value of

ΔE (energy converted per cycle) given in this table is $1/N \, \Sigma_0^N \Delta E$. In addition to the specimens Nos. 1 to 9, for which the complete observations during test are given in figures 3 and 4, the end points of the test (corresponding to fatigue failure) are given in these figures for four more specimens, Nos. 10 to 13. Three of these specimens (Nos. 10, 11 and 13) failed prematurely but are included to indicate the lower limits of endurance observed in tests on this alloy.

The dependence of damping capacity on strain after a given number of cycles and for particular values of the alternating couple may be obtained from figures 3 and 4. For example, this dependence is shown in figure 6 at 0·25, 1·0, 2·5, 10, 25 and 100 million cycles.

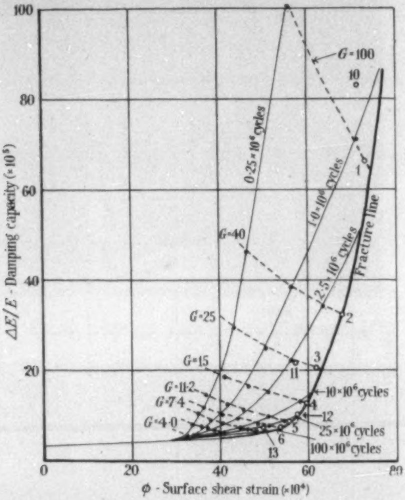


Figure 6. Variation of damping capacity with strain for an alloy of aluminium containing 11% Mg.

It is evident that changes occur very rapidly in the material during the first million cycles, especially for high values of the applied couple (G). The bold curve connects points corresponding to the occurrence of fracture in specimens Nos. 1 to 6.

Alloy containing 5% *magnesium*

The general behaviour of the alloy containing 5% magnesium is similar to that of the 11% alloy and it is not proposed to record detailed observations as in figures 3 and 4. From data similar to those given in these figures, the variation damping capacity with strain for 0·25 million cycles and at fatigue failure has been derived and is shown in figure 7 together with the corresponding graphs taken from figure 6 for the 11% alloy.

It is evident that the strains which will cause fatigue failure in the 5% alloy are appreciably less than the strains necessary to cause failure of the 11% alloy, the shaded regions in figure 6 showing the limiting conditions of damping capacity and vibrational strain within which fatigue failures have been observed. The figures adjacent to these regions indicate the number of cycles required to cause fatigue failure under conditions of constant peak value of the applied couple. Comparison of these " fatigue regions " with the corresponding curves representing the alloys after vibration for 0·25 million cycles shows that pronounced changes in physical condition occur before failure by fatigue.

Other alloys

Although this paper deals mainly with observations on two alloys, it may not be inappropriate to record that the method of test is applicable to some other aluminium alloys. Figure 8 shows specimens of several alloys in which fatigue cracks have been produced, although no quantitative examination of their fatigue properties has yet been attempted by this method. The specimens shown in figure 8 had natural frequencies of about 1500 cycles per second, and after the completion of the test were anodized to show the fatigue cracks.

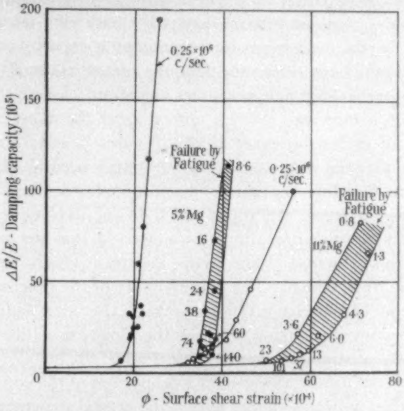


Figure 7. Variation of damping capacity with strain for alloys of aluminium containing 5% and 11% Mg.

§7. DISCUSSION OF RESULTS

For a particular deformation, the damping capacity of the material may, with certain reservations, be regarded as a measure of the ratio of the total plastic strain per cycle to the maximum elastic strain.*

---

* Strictly this may be assumed, for torsional deformation of a solid cylinder, only when the damping capacity is independent of strain. Under such conditions $\Delta E/E=2\Delta\phi/\phi$, where $\Delta\phi$ is the plastic component of the total strain ϕ. When the damping capacity is not independent of strain, the measured value of the energy dissipated corresponds to an integration of the energies dissipated at strains progressively increasing from zero at the vibration axis to the measured value of ϕ at the surface. For the purpose of discussion it is proposed to assume that $\Delta\phi$ as defined in the above equation represents an average plastic strain for a maximum surface strain ϕ.

Experimental evidence suggests that all materials have a finite, but often very low, damping capacity at very low strains; for the alloys under investigation this small and practically constant damping capacity is estimated to be 3×10^{-5} for surface shear strains up to 30×10^{-4} (5·4 tons per sq. in.) for the 11% alloy and 15×10^{-4} (2·7 tons per sq. in) for the 5% alloy. The occurrence of a small damping capacity at even the lowest strains does not necessarily indicate that purely elastic strains are non-existent; C. Zener (1940) has shown that a small damping capacity may arise from causes other than plastic deformation, although under the conditions of the present investigation these phenomena are not likely to contribute to the observed damping capacity. It is possible that the practically constant damping capacity for small strains may be due to localized slip of a type which may be considered as reversible in the sense that no damage is caused to the crystal structure. The centres of dissipation may be imperfections of the type suggested by G. I. Taylor (1934) and others, where small groups of atoms may be able to assume alternative positions without materially modifying the general structure. Whatever may be the mechanism causing damping capacity at small strains, it is evident that in the range where the damping capacity is small and constant, the alloy shows proportionality between stress and strain.

As the strain is increased above a critical value the damping capacity (for a small number of cycles) increases rapidly (figures 6 and 7) and consequently proportionality between stress and strain no longer occurs; strains of 15×10^{-4} and 30×10^{-4} are, therefore, the limits of proportionality for the 5% and 11% alloys respectively under dynamic conditions which do not involve prolonged vibration. Above this limit of proportionality it is suggested that the mechanism of the dissipation of energy is true plastic strain, i.e. strain which changes the original crystal structure irreversibly. The irreversibility of the process is demonstrated by the dependence of the relations (figures 6 and 7) between damping capacity and strain on the amount of vibration to which the alloys are subjected.

Consequently, at strains above the critical value, the metal may be expected to change in properties during vibration if a proportion of the converted vibrational energy is absorbed as strain energy within the crystal structure. Absorption of energy would result in strain hardening of the metal, and this expectation is confirmed, for figures 4, 6 and 7 show that the damping capacity decreases during continued vibration at strains above the limit of proportionality. Under suitable conditions of vibration the limit of proportionality may be increased from 15×10^{-4} to 34×10^{-4} (2·7 to 6·1 tons per sq. in. shear stress) for the 5% alloy and from 30×10^{-4} to 50×10^{-4} (5·4 to 9·0 tons per sq. in.) for the 11% alloy.

Since strains above the primary limit of proportionality cause changes in the metal it is to be expected that this will be a limiting strain below which fatigue failures will not occur. Figure 7 shows, in fact, that all specimens which failed by fatigue did so at strains greater than the secondary limit of proportionality, i.e. the limit of proportionality after prolonged vibration. The values of damping capacity and strain at which the fatigue failures occurred all lie for each alloy upon a curve which is very similar in shape to that for the damping capacity and strain after a small amount of vibration, but displaced in the direction of higher strain.

It is proposed now to consider the factors which may determine the number of cycles required to cause a fatigue failure. It has been indicated that failure of these alloys is preceded by strain hardening. If $\Delta\phi$ is taken* to represent the amount of plastic strain per cycle (including also the slip corresponding to the damping capacity in the strain range where the damping capacity is constant but small), an expression for the total equivalent plastic strain at fracture is $\Sigma_0^N \Delta\phi$. It is suggested that fatigue failure may depend on this quantity. Another possibility is that fatigue failure is related to the total amount of energy dissipated up to the time of fracture. The total energy dissipated (excluding external losses) is given in table 1 for several specimens of the 11% alloy. Referring at present particularly to this alloy, figures 9 *a* and 9 *b* show how $\Sigma_0^N \Delta\phi$ and $\Sigma_0^N \Delta E$ vary with N, the total number of cycles to cause fracture (or 10^8 cycles if fracture does not occur before then). The gradient of the line OP represents the average plastic strain per cycle whilst the gradient of the line OQ represents the average dissipation of energy per cycle for specimen No. 7, i.e. the specimen subjected to the highest value of the alternating couple among those specimens which were not fractured after 10^8 cycles. It will be seen that the points corresponding to the fractured specimens (Nos. 1 to 6) lie approximately on a straight line above and roughly parallel to OP or OQ, whilst the points for specimens 8 and 9 lie below these lines.

The criterion for failure by fatigue may be either, from figure 9 *a*,

$$\Sigma_0^N \Delta\phi = A + BN \qquad \ldots\ldots(3)$$

or, from figure 9 *b*,

$$\Sigma_0^N \Delta E = C + DN, \qquad \ldots\ldots(4)$$

where A, B, C and D are constants. The scatter † of the points representing fatigue failures makes it difficult to decide which of these equations is most likely to represent the truth, but the scatter is somewhat less in figure 9 *a* than in figure 9 *b*.

In equation (3), the constant B is the amount of slip per cycle which occurs without contributing to fatigue failure of the specimen, i.e. it represents that part of the mechanism causing damping capacity, which exists at all strains and is predominant before true plastic strain occurs. The constant A is the total amount of irreversible strain occurring before failure by fatigue. The constants C and D may be interpreted similarly in terms of energy.

Thus it appears probable that fatigue failures occur either after a certain amount of irreversible strain or, perhaps less definitely, after a certain conversion of energy associated with plastic strain.

It has been stated already that the experimental results suggest that failure occurs after a definite amount of strain hardening, and this would be expected to occur after a definite amount of irreversible strain, but not necessarily after a definite conversion of vibrational energy, since a considerable proportion of the energy represented by the damping capacity may appear as thermal energy. The energy absorbed as strain energy in the crystal structure is probably roughly proportional to the conversion of energy associated with plastic strain.

---

* It should be noted that under the conditions of operation of the vibration equipment (i.e. constant exciter current) $\Delta\phi$ is practically constant (it would be strictly constant if no external loss of energy occurred) throughout the experiment and is proportional to G, the peak value of the applied alternating couple. It is therefore independent of the amount of strain hardening which may occur during the test.

† This scatter is to be expected in fatigue tests involving a number of specimens because the formation of fatigue cracks is influenced by the conditions of the surface of the metal. The points representing fatigue failure of specimens 10 to 14 of this alloy are also included in figures 9 *a* and 9 *b*, and it is seen that these also lie above the lines OP and OQ.

The experimental data for the 5% alloy may be interpreted similarly to that of the 11% alloy and it is only necessary to indicate differences in the magnitude of the effects which occur.

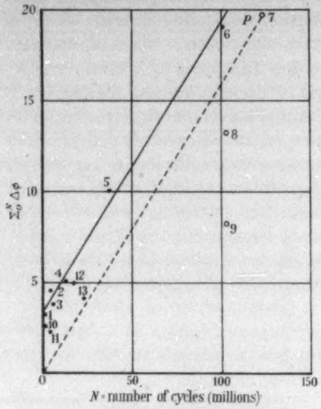


Figure 9*a*. Total "plastic" strain before fracture of aluminium alloy containing 11% Mg.

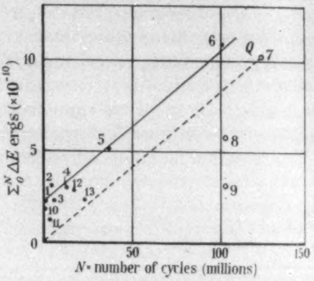


Figure 9*b*. Total energy converted before fracture of aluminium alloy containing 11% Mg.

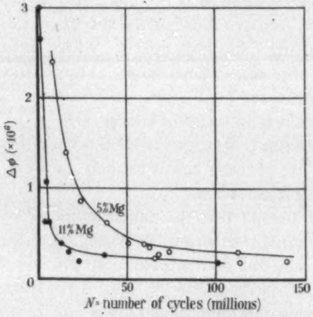


Figure 10. Variation of endurance with plastic strain per cycle.

It is apparent that the endurance of these alloys is dependent on the amount of plastic strain per cycle and the difference between the two alloys is well illustrated by graphs showing endurance as a function of $\Delta\phi$, the plastic strain per cycle, this quantity remaining constant for any one specimen under the present conditions of test. The relation between endurance and $\Delta\phi$ is shown in figure 10 for the two alloys. The curves are roughly hyperbolic, in agreement with equation (3), and indicate that the constant B (slip per cycle not contributing to fatigue failure) is probably the same for both alloys and of the order of $1 \cdot 5 \times 10^{-7}$. The constant A, which represents the total amount of irreversible strain developed before failure by fatigue, is estimated to be 5·8 times as great for the 5% alloy as for the 11% alloy.

§8. CONCLUSIONS

Fatigue failure of aluminium alloys containing 5% and 11% of magnesium is preceded by strain hardening. Strain hardening commences when the shear strain exceeds the primary limit of proportionality of the alloy. Below this strain the damping capacity, which is small (3×10^{-5}) and substantially constant, is not associated with plastic strain as normally understood, but, above this strain, work done in causing plastic strain provides an increasingly important contribution to the damping capacity.

For strains exceeding the primary limit of proportionality, the damping capacity is a function of time of vibration, decreasing as the metal strain hardens to an amount ultimately equivalent to an increase of the limit of proportionality by about 120% for the 5% alloy and about 65% for the 11% alloy. The results suggest that fatigue failures occur when the total amount of irreversible strain caused by vibration reaches a certain value characteristic of the alloy, endurance being dependent on the amount of irreversible strain per cycle.

§9. ACKNOWLEDGMENT

The author wishes to thank the Directors of High Duty Alloys Ltd. for permission to publish this paper.

REFERENCES

Hanstock, R. F. and Murray, J., 1946. *J. Inst. Met.*, **72**, 97.
Taylor, G. I., 1934. *Proc. Roy. Soc.*, A, **145**, 362.
Zener, C., 1940. *Proc. Phys. Soc.*, **52**, 152.

REFRACTION EFFECTS IN ELECTRON DIFFRACTION

By J. M. COWLEY and A. L. G. REES,
Division of Industrial Chemistry, Council for
Scientific and Industrial Research, Melbourne

MS. received 8 *October* 1946

ABSTRACT. Electron diffraction rings from MgO and CdO smokes have previously been shown under high resolution to be multiple. More detailed observations of this phenomenon have been made. It has been observed that the arcs given by tilting of oriented specimens are divided into several components, which vary in intensity and separation with angle of tilt of the specimen. Spotty rings given by brown CdO (containing excess metal atoms) and some MgO specimens are composed of short streaks forming groups of up to six radiating from the expected position of a single-crystal reflection, and in some cases these streaks are divided into inner and outer components.

The theory that the fine structure of the rings arises from refraction at the faces of the regularly shaped particles (cubes) was suggested by Sturkey and Frevel (1945). On the basis of this theory, general expressions have been derived for the deviation of the direction of the diffracted beam owing to refraction at the faces of cubes. Application of these expressions to particular cases has shown complete agreement between the calculated and observed form of the fine structure of the rings, arcs and groups of streaks. Quantitative agreement for the separation of the arc components is obtained by assuming values of inner potential of the order to be expected. The value of the inner potential varies with the reflecting plane from 12 to 16 volts. For the doubled streaks the elongation of the inner component represents a variation of inner potential about this value, arising from excess metal atoms in the lattice. The outer component represents an unexplained value of about 25 volts.

Streaks observed on ZnO smoke rings are explained as due to refraction by the long thin spines characteristic of this material. It is shown that refraction will have the effect of broadening diffraction rings, even for irregular particles, and will give rise to line breadth and intensity anomalies in diffraction patterns from specimens with regular crystal habit.

---

§1. INTRODUCTION

The development of high resolution electron-diffraction cameras has resulted in the observation of fine structure not previously suspected in electron-diffraction patterns. Hillier and Baker (1945), for example, noted a multiplicity of the rings in transmission patterns from MgO smoke particles; they found the rings 111, 222, 224 and 226 to be double, 200 single, and the 220 ring to possess an unusual contour, suggesting five components. Sturkey and Frevel (1945), in similar experiments on MgO and CdO, made the generalization that *hhh* reflections were invariably double, *h*00 single and others broadened. Sturkey and Frevel attributed this multiplicity to refraction by particles of regular geometrical shape (electron micrographs show them to be perfect cubes), arising from the inner potential (mean potential above free space) within the crystal. This appeared to be confirmed by the variation in the intensity of the reflection from 220 planes observed with change of accelerating potential. By use of a geometrical device referred to as a " triangle of reflection " formed by the intercept of the plane of reflection on the reflecting plane and the cube faces, they derived an expression relating the angular displacement of the components δ to the angles of refraction r_1 and r_2 at the sides of this triangle, viz.:

$$\delta = \frac{P}{2E}(\pm \tan r_1 \pm \tan r_2),$$

where P is the inner potential in volts and E the electron accelerating potential. From this expression P was calculated to be 12 ± 4 volts for the 111 doublet. It will be shown in §4 of this paper that this expression is a special case of a more general expression and is not generally valid.

In the work described here (see also, Cowley and Rees (1946)) more detailed observations of this phenomenon have been made and several new features of fine structure have been discovered. The theory of refraction by particles of regular shape has been developed fully and has been shown to account quantitatively for the observed features of these patterns. Previously, refraction effects have been considered negligible at the accelerating potentials employed (50 kv.) for all but small reflection angles, but they assume a greater importance in high resolution electron diffraction. The refraction effect is shown to make large contributions to line breadth and, except for particles of diameter <300 A., it represents the predominating factor. Moreover, the unexplained anomalous intensity distributions in certain electron-diffraction patterns are explained quite naturally on the basis of refraction by particles approaching regular geometric forms.

§ 2. EXPERIMENTAL

The electron diffraction patterns were taken with an R.C.A. type EMU electron microscope used with the diffraction adaptor. The precise electron-optical system and the very stable H.T. and lens supplies endow the instrument with high resolution characteristics. Resolution in focused patterns was found to be 12μ or better. Since the system employs an electromagnetic lens between the specimen and the photographic plate to focus the pattern, calibration is necessary to obtain the effective specimen-plate distance. Pure, recrystallized sodium chloride was employed as a standard.

Examination and measurement of fine structural details were made on enlargements to 50 diameters, attained in two stages. Some of the enlargements on which these measurements were made are reproduced in the plate.

The materials studied, namely, MgO, CdO and ZnO, were mainly prepared by the oxidation of the corresponding metal vapour in air. The oxide smoke was collected either directly on a 200-mesh stainless steel gauze or on a thin collodion film (<200 A. thick) supported across the gauze. Materials other than smokes were prepared for examination by the evaporation of suspensions in water or secondary butyl alcohol directly on the collodion film. CdO formed at some distance from the evaporating metal was the usual yellow-orange colour ; that formed close to the metal was dark brown, since conditions are here more favourable for incomplete oxidation with the formation of a non-stoichiometric oxide containing an excess of cadmium in the oxide phase. As it happens, these two forms of oxide exhibit significant differences in the fine structure of the diffraction patterns which will be discussed later.

Data on particle size and shape were obtained from electron micrographs. Both MgO and CdO smokes consist of perfect cubic particles, and ZnO smoke of fine needles arranged in "fourlings" together with occasional semi-transparent plates. For MgO and yellow CdO the average cube edge was ~500 A. ; for brown CdO the particles were somewhat larger (cube edge ~1000 A.) and there were present some sintered aggregates of small numbers of particles. In ZnO smoke, it is shown that the [0001] axis is parallel to the axis of the spikes and it is probable, although difficult to establish by direct observation, that the spikes have hexagonal cross-section.

§ 3. RESULTS

The results will be discussed in relation to three types of diffraction pattern observed, viz. :

(i) Continuous ring patterns showing subdivision of the normal reflections.

(ii) Arc patterns obtained from specimens possessing a preferred orientation.

(iii) Ring patterns showing the individual single crystal reflections as groups of spots or streaks, which occurred when only a small number of crystals were present in the irradiated area of the specimen.

Continuous ring patterns

Patterns given by MgO and CdO smokes showed a subdivision similar to that previously reported. Our observations, tabulated in table 1, are more detailed than those referred to.

Table 1

Reflection	Multiplicity	Remarks
*h*00	Single	Sharp
hhh	Doublet	Clearly resolved
*hh*0	Triplet	Central component sometimes very weak, resembling doublet
*hk*0	Triplet	——
331 422 531	Doublet	——
331, 511, 442, 622, 642	Triplet	——

The two components of a doublet appear one on either side of the normal position. In triplets, the central component is undisplaced and the outer components displaced symmetrically. In figure 1 (*a*) an enlargement of a portion of one of these patterns showing resolution of these components is given.

Arc patterns

In many specimens supported on collodion films, a large majority of the cubes were oriented with the cube face parallel to the film, so that, on tilting the film away from the position normal to the beam, a pattern of arcs lying on layer-lines corresponding to $l=0$, ± 1, ± 2 was observed for *hkl* planes recorded. These arcs showed multiplicity and, in view of the greater intensity in the arcs, more information could be obtained regarding the number and separation of components, since the weaker components were more evident. The intensities and separations of the components varied with the angle of tilt and were, in fact, different for the different arcs on the one ring.

Component separations, measured on a travelling microscope, and visually estimated intensities, are given for both rings and arcs for CdO in table 2. A selection of the data only is recorded, since the complete set of data would be too extensive to reproduce here. Data have been obtained for arcs on CdO patterns at seven angles of tilt between 0° and 60°. The components are represented by letters indicating their relative intensity (strong (S), medium (M) and weak (W)) and the letters are separated by numbers indicating the separation of these components in units of 10^{-5} radian. Thus the record, W42S44S42W, indicates that two strong components occur at positions $\pm 22\times 10^{-5}$ rad. from the expected position and that two weaker components occur outside the strong component at $\pm 64\times 10^{5}$ rad. displacement. It will be noted that there is a marked change in the intensities and separations of components of any one arc with increased

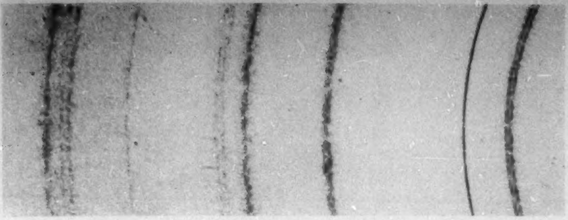


Figure 1 (*a*). Electron diffraction pattern from yellow cadmium oxide showing resolution of ring components. Enlargement 16 diameters.

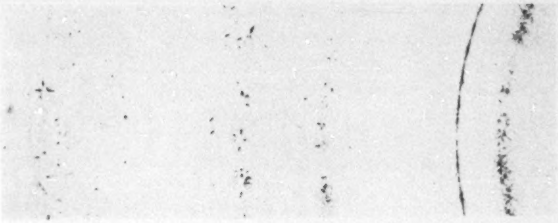


Figure 1 (*b*). Electron diffraction pattern from brown cadmium oxide (excess cadmium) showing resolution of single crystal reflections. Enlargement 16 diameters.

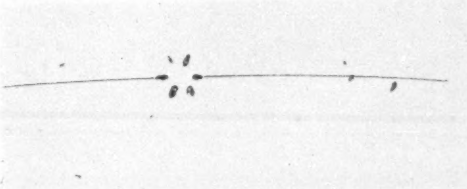


Figure 2. Group of six streaks on 422 ring arising from single crystal of magnesium oxide. Enlargement 56 diameters. Undisplaced ring position indicated by continuous line.

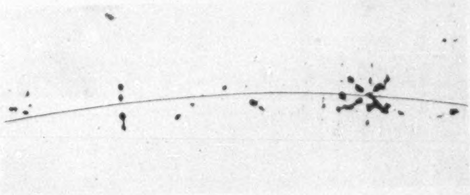


Figure 3. Groups of streaks on 220 ring arising from single crystals of magnesium oxide. The second component on each streak is clearly visible. Enlargement 56 diameters. Undisplaced ring position indicated by continuous line.

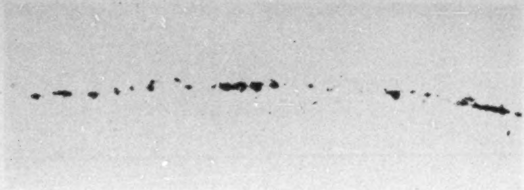


Figure 4. Single crystal reflections along the 200 ring. Enlargement 56 diameters.

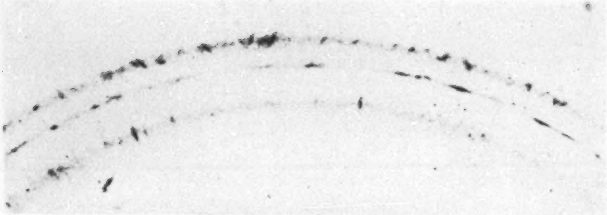


Figure 12. Portion of diffraction patterns from zinc oxide smoke particles showing streaks on 100, 002, 101 rings. Enlargement 24 diameters.

Figure 13. Portion of diffraction patterns showing 222, 400, 420 and 422 rings from near cubic particles. Enlargement 24 diameters.

angle of tilt. Examples of this behaviour are shown in figures 7, 8 and 9, in which the theoretical relation between separation and relative intensity and angle of tilt are compared with the experimental results.

Patterns showing single crystal reflections

Diffraction patterns from specimens having a relatively small number of individual crystals in the illuminated area exhibit "spotty" rings and the spots appear to be grouped together in groups of two to six, where the spots lie on the various ring components. This grouping was more easily observed in patterns from brown CdO and MgO collected close to the metal surface, in which the spots were elongated into short streaks radiating for each group from a common point lying on the normal ring position (figure 1 (*b*)). Groups of up to six streaks were observed with the undisplaced component of a group of spots usually absent. This phenomenon is illustrated in figures 2 and 3. There is a noticeable tendency for these streaks to make angles with the radius of the pattern which are characteristic of the ring in question. For example, for both MgO and CdO the angles are (i) 45° for 111, (ii) 0° or 45° for 220, (iii) 90° for 200, etc. For the 200 reflection, the two components lie along the ring, which therefore appears single, a fact which explains the relative sharpness of this ring. This is illustrated in figure 4. For patterns showing streaks and groups of streaks, the angles between the streaks and the radius of the ring were measured for various rings on 50-diameter enlargements. Definite maxima appeared in the frequency plots, confirming the existence of preferred angles. Maxima were found in such plots for CdO at 30° to 40° for the 222 ring, at 0° and 50° for 220 and at 15°, 30° and 70° for 311. One of the plots is reproduced in figure 10 accompanying the theoretical discussion of these results in § 4.

Another feature of certain of these patterns, particularly evident in the inner rings, is the presence of a second component of each streak in a group; this is illustrated in figure 3. It was found that the ratio of distances from the centre of a group to the outer and inner components was constant for a given ring. This observation is discussed in § 7.

§ 4. THEORETICAL

Many phenomena observed in X-ray and electron diffraction, which bear some resemblance to the effects described, suggest interpretations different from that developed later in § 4. It was considered desirable to examine these effects and to show where these interpretations failed.

In many single crystal X-ray patterns, "extra" spots, usually temperature dependent and diffuse, are observed and these have been satisfactorily explained by the Faxén–Waller theory as arising from thermal vibration of the atoms comprising the crystal. Most reminiscent of the effects described in § 3 are the secondary reflections observed by Lonsdale (1945) in type 1 diamonds; these spots are sharp and not markedly temperature sensitive. W. H. Bragg (1942) has interpreted these extra spots on the basis of extended reciprocal lattice points in small crystals. Similarly, v. Laue (1936) has explained the occurrence of groups of spots in electron-diffraction patterns as arising from the shape of the crystals. He pointed out that each reciprocal lattice point would possess horns or extensions

Table 2

Ring	Ring profile	l	$\alpha=10°$	27°	29°	40°	46°	52°	60°
111	W51S41S51W	1				W42S44S42W	S41S	S24S	W46S41S46W
220	W44S24S44W	0	W61S61W	S44M44S	S60S	S43S	W55S46S55W	S37S	W49S35S49W
		2					W47S31S47W	W46S34S46W	W57S52S57W
311	W44S44W	1		S41S	W39S29S39W	W49S49W	W46S46W	W43S43W	W44S21S44W
420	W42S42W	0	W46S46W	S44S	W51S32S51W	W46M28M46W	W44M22M44W	W41M19M41W	W35S17S35W
		2		M46M46M	W44M44W	W43M43W	W38M20M38W	M23M	M17M
440	—	0	W62M62W	M41W41M	M62M	W60W	W62W	W51W	W43W

normal to the bounding faces of crystals and that these would be more pronounced for small crystals of regular shape. On this basis v. Laue was able to explain the results of Cochrane (1936) and Brück (1936) for layers of Ni, Co and Ag on various substrates.

Whilst these effects superficially resembled those reported here and the order of crystal size involved in both cases was the same, they differed in several significant respects. All these apparently related effects are of a different order of magnitude, and in every case the central component of observed groups of spots would necessarily be strong, whereas in the single crystal groups observed in our patterns the central spot is always absent. Moreover, on v. Laue's theory, the extensions associated with each reciprocal lattice point for cubes would be symmetrical and independent of orientation of the crystal, whereas the structures of arcs on the one ring have been shown to differ widely and also to vary with orientation. Predictions based on this theory regarding the subdivision of rings, the distribution of angles between streaks and radius, and the angles between streaks and radius in individual groups of spots, did not agree with the observations recorded in § 3.

Daniel and Lipson (1943) observed side bands in x-ray powder patterns from the phase Cu_4FeNi_3, attributable to the presence of a superlattice formed by the regular segregation of different atoms. However, a strong central component is expected here also. An interpretation on this basis would require $h00$ rings to be doubled, whereas our observations show them to be single.

Rooksby (1943) has reported similar multiplicity in x-ray powder patterns from NiO. $h00$ rings were observed to be single, whilst other reflections were either double or triple. Rooksby was able to account for these observations by assuming that the structure was only pseudo-cubic and was, in fact, rhombohedral. This interpretation cannot be used for the results of this investigation since it does not account for the grouping of spots or streaks; moreover, spots would be undivided and on one component or the other.

X-ray powder patterns of the MgO and CdO used, taken with a 14 cm. camera, exhibited no anomalies of the type found in electron diffraction.

Refraction theory

The explanation of the fine structure features on the basis of reflection suggested by Sturkey and Frevel (1945) has been investigated more thoroughly and found to predict the observed effects accurately. The vector method of Luneberg (1944) has been applied to the general case.

For reflection or refraction at a plane surface L_i (figure 5),

$$\overline{S}_{i+1} = \overline{S}_i + \Gamma_i \,.\, \overline{M}_i.$$

$\overline{S}_i = n_i \overline{T}_i$, where n_i is the refractive index and $\overline{T}_i$ the unit vector along the beam. Γ_i is a scalar function of the angle of incidence ϕ_i and $\overline{M}_i$ is the vector perpendicular to the plane L_i. Putting $\rho_i = n_i \cos\phi_i = \overline{S}_i \,.\, \overline{M}_i$, we have $\Gamma_i(\rho_i) = -2\rho_i$ for reflection and $\Gamma_i(\rho_i) = (n_{i+1}^2 - n_i^2 + \rho_i^2)^{\frac{1}{2}} - \rho_i$ for refraction.

For a ray passing through p faces

$$\overline{S}_{p+1} = \overline{S}_1 + \sum_{i=1}^{p} \Gamma_i(\rho_i) \,.\, \overline{M}_i.$$

For a ray passing through two faces of a cube, with normals $\bar{M}_1$, $\bar{M}_2$, and being reflected internally at the Bragg angle θ at a plane with normal $\bar{N}$ (figure 6), we may write

$$\bar{S}_4 = \bar{S}_1 + \Gamma_1(\rho_1)\bar{M}_1 + \Gamma_N(\rho_N)\bar{N} + \Gamma_2(\rho_2)\bar{M}_2,$$

which may be reduced to

$$\bar{S}_4 = \bar{S}_1 - 2n\sin\theta\,.\,\bar{N} + \frac{n^2-1}{2\cos\phi_1}\,.\,\bar{M}_1 - \frac{n^2-1}{2n\cos\phi_2}\,.\,\bar{M}_2.$$

Since $n^2-1=P/E$ for an inner potential of P volts and an accelerating voltage of the beam E, and n is close to unity, we have, taking account of the possible variation in signs,

$$\bar{S}_4 = \bar{S}_1 - 2\sin\theta\,.\,\bar{N} + \frac{P}{2E}\left(\pm\frac{\bar{M}_1}{\cos\phi_1} \pm \frac{\bar{M}_2}{\cos\phi_2}\right).$$

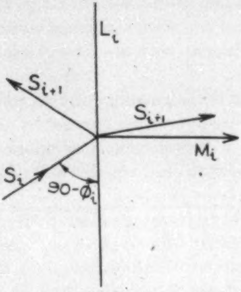


Figure 5.

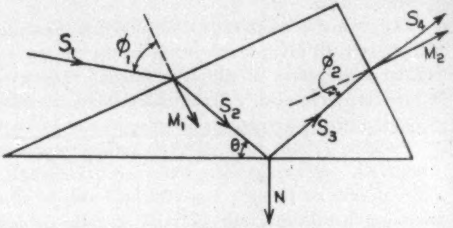


Figure 6.

Here the second term represents the deviation due to Bragg reflection, and the third term represents the added deviation due to refraction. The direction of the refraction displacement will be governed by the direction of the surface normals. For cubes there are, in general, six pairs of non-parallel faces through which the beam may pass, corresponding to various combinations of the face normals $\pm\bar{M}_1$, $\pm\bar{M}_2$, $\pm\bar{M}_3$. Hence there may be deviations in six directions, giving rise to a group of up to six spots surrounding the normal spot position, or, in the case of varying inner potential, six streaks radiating from the common centre. Pairs of parallel faces give no displacement. The undisplaced spot is usually weak and often missing owing to absorption in passing through the whole length of the cube. This theory then gives at least a qualitative explanation of the groups of spots and streaks.

Arc patterns

In the patterns of arcs given by oriented cubic crystals on tilted films, each arc is given by many cubes in almost the same orientation with respect to the beam, and hence the arc will be generated by translating a certain spot-group along the circumference of the ring. Thus the arcs will be multiple and the distribution of components across the arc will be given by the projection of the appropriate group of spots on the radius, that is, by projecting the deviations δ upon the direction of N, the normal to the diffracting plane.

Suppose the beam makes angles ϕ_1, ϕ_2, ϕ_3 with the normals $\bar{M}_1$, $\bar{M}_2$, $\bar{M}_3$, and these normals make angles ψ_1, ψ_2, ψ_3 with $\bar{N}$. Assuming the Bragg angle θ and the angular deviation δ to be negligibly small compared with ϕ_1, ϕ_2 and ϕ_3, the projections of the deviations δ on $\bar{N}$ are of the form

$$\delta_N = \frac{P}{2E}\left(\pm\frac{\cos\psi_1}{\cos\phi_1} \pm \frac{\cos\psi_2}{\cos\phi_2}\right).$$

In the special case where the path of the beam lies in a plane perpendicular to a cube edge, this becomes

$$\delta_N = \frac{P}{2E}\,(\pm\tan\phi_1 \pm \tan\phi_2),$$

which is the form given by Sturkey and Frevel (1945).

In general, this degeneration is not valid, since the path of the beam is not normal to the cube edge.

In an arc there will be a maximum of seven components including the component with zero displacement; equal indices or zero indices of the diffracting plane will reduce this number. The relative intensities of the components of an arc may be calculated from consideration of the areas presented to the beam corresponding to each component. Two limiting cases may be distinguished—(i) if absorption is neglected the whole area of a cube presented to the beam is effective; (ii) for very high absorption only a thin strip along the cube edges will contribute and the width of the strip will vary with the angles involved.

For the (220) *plane*, for example, the angles between face normals and $\bar{N}$ are $\psi_1 = 90°$, $\psi_2 = \psi_3 = 45°$ and

$$\phi_2 = \phi_3; \qquad \cos\phi_2 = 1/\sqrt{2}\sin\phi_1.$$

The possible values of $(\delta_N \,.\, 2E/P)$ will then be

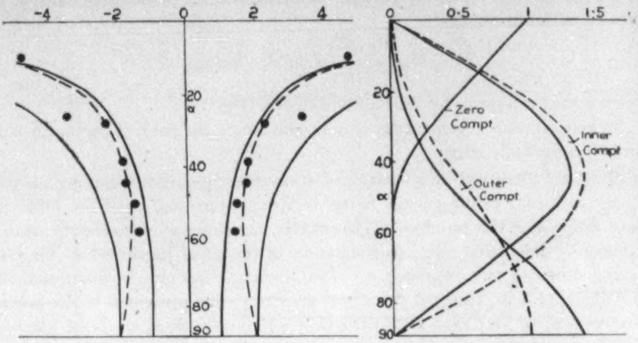
$$0, \quad \pm\frac{1}{\sin\phi_1}, \quad \pm\frac{2}{\sin\phi_2}.$$

Then ϕ_1 will be the angle of tilt α of the specimen with respect to the beam, and so the displacement of the components of the 220 arc may be plotted against α to give the curves of figure 7 (*a*). The relative intensities of the components are calculated readily for the two limiting cases of zero and high absorption. With the exception that in the latter case the zero displacement component is absent, there is no great difference in the relative intensities, as is shown in figure 7 (*b*). In agreement with observations (table 2) the zero displacement component decreases in intensity with angle of tilt. In practice the two displaced components are not resolved. The dotted curve of figure 7 (*a*) represents a mean displacement weighted for intensities. This is plotted for the case of no absorption, which differs very little from that for high absorption.

The measured displacements, δ_N, from CdO patterns for various angles of tilt are plotted as full circles in figure 7 (*a*), with a multiplying factor of $2E/P$, where P is chosen as 12·7 volts to give the best agreement. The value of the inner potential for CdO on the assumption of no absorption is thus 12·7 ± 2·0 volts, and for high absorption it is 12·3 ± 2·0 volts.

Measured values of the displacements for the 440 arc fitted the same curve for $P = 15 \pm 2$ volts.

Corresponding calculations for the 202 arc on the $l = 2$ layer line are made by taking the angle of tilt $\alpha = \phi_2$. This arc does not appear until $\alpha = 45°$, as is shown in the graph of displacement against α in figure 8. The relative intensities are indicated approximately by the thickness of the curves. Similarly, the displacements for the $l = 0$ and $l = 2$ arcs on the 420 ring have been plotted in figure 9.



(a) (b)

Figure 7 (a). Plot of calculated displacement of the components of the 220 arc against angle of tilt. Experimental data—full circles.
(b). Plot of relative intensity of components of the 220 arc against angle of tilt. Continuous line—zero absorption. Broken line—high absorption.

The outer two components will not be resolved and a curve is therefore drawn representing the mean. The plotted points are for CdO arcs with $P = 16{\cdot}7$ volts for $l = 0$ and $P = 15{\cdot}6$ for $l = 2$. Similar calculations have been made for 111 and 620 arcs. Values of inner potential found for the various arcs are tabulated in table 3. Those for which the accuracy is not indicated are less accurate than the others.

Table 3

Arc		CdO	MgO
111	$l=1$	11	13
220	$l=0$	12·7±2·0	—
	$l=2$	10	—
222	$l=2$	14	17
420	$l=0$	16·7±1·5	15
	$l=2$	16	15
440	$l=0$	15·0±2·0	15
	$l=4$	13	10
620	$l=0$	15	11
	$l=2$	17	12

It will be noticed that the values of P for the second orders of 111 and 220 arcs are higher than for first orders. The difference is within the possible experimental error, but is sufficiently consistent to indicate a real difference. The error introduced in the calculations by assuming the Bragg angle θ to be negligible is in the right direction, but further calculations have shown that the magnitude of the error introduced by this assumption is not sufficiently large (less than about 3%, except for very large displacements of very weak components of the arcs) to explain this difference.

The variation of the value of inner potential between different reflections is greater than the probable experimental error, and indicates that there is a dependence of the effective inner potential on the direction of the electron path through the crystal. The values of the inner potentials obtained for CdO and MgO are of the order expected, but, apart from the value 12 ± 4 volts given by Sturkey and Frevel (1945), there are no values given in the literature with which these results may be compared. However, the qualitative and quantitative agreement of the observed forms of the arcs and those calculated on the basis of the refraction theory is complete within experimental error, provided these values of P are taken as correct.

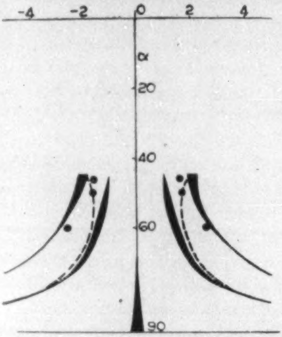


Figure 8. Plot of calculated displacement of components of 202 arc against angle of tilt. Breadth of curve gives indication of dependence of relative intensity on tilt. Experimental data—full circles.

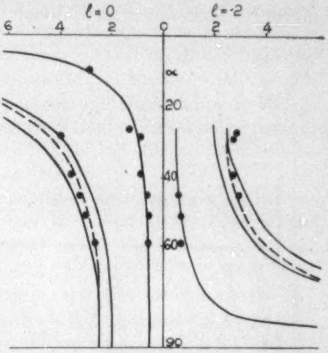


Figure 9. Plot of calculated displacement of components of 420 arc, for both $l=0$ and $l=2$, against angle of tilt. Experimental data—full circles.

Patterns showing streaks and groups of streaks

The form of a group of streaks for any given orientation of a cube giving a particular reflection may be predicted from the refraction theory. Without detailed calculations it is possible to derive general conditions satisfied by the angles between the streaks of a group and the radius for any particular ring. The three pairs of diametrically opposite streaks will define three angles with the radius. For the 200 ring there will be only one pair of streaks and the corresponding angle will be 90°, i.e. the streaks lie along the ring; for 220, one angle will be 0°, the other two will be equal and lie in the range 45–90°; and so on.

Many measurements were made of the angles between the streaks of individual groups and the radius, and in every case the angles complied with the above conditions within the limits of experimental error.

Calculations of the distribution of the angles between streak and radius were made for the 220 ring. The probability of a given angle between streak and radius occurring was calculated on the assumption that all angles of tilt α are equally probable. This probability was multiplied by the intensity of the streak to provide a measure of the chance of that particular angle being observed. This product has been plotted against angle in figure 10. The number of angles measured in ranges of 4° is plotted in the same figure as a block-diagram. Allowing for the inaccuracy of measurement of angles the agreement is satisfactory.

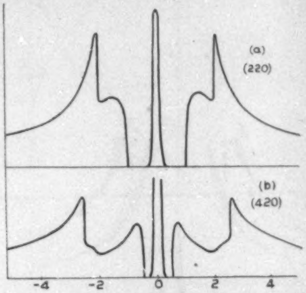


Figure 11. Calculated intensity profile for (*a*) 220 ring, (*b*) 420 ring. Relative intensity of central components will vary markedly with crystal size.

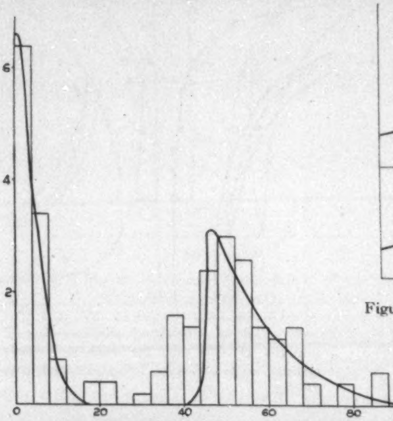


Figure 10. Calculated distribution of angles made by individual streaks with radius of diffraction ring for 220 reflections. Experimental data plotted as block diagram.

§ 5. RING PROFILES AND RING BREADTH

Cubic crystals

Rings given by random cubes may be considered as the sum of the arc patterns for all angles of tilt in all positions.

From the method outlined in the discussion of arc patterns in § 4 for obtaining the displacement and intensity of each component in terms of the angle of tilt, the intensity of each component may be expressed as a single-valued function of displacement. Summing the intensities due to the several components for each displacement gives the intensity profile of the continuous ring. This has been done for the 220 and 420 rings, and the profiles obtained on the assumption of no absorption are shown in figures 11 (*a*) and 11 (*b*). The profiles calculated for the case of high absorption are closely similar, except for the absence of the central component. In practice the broadening of each part of the profile due to finite crystal size and finite beam diameter will permit the resolution of only the gross features of the contours. Thus the 220 ring is a doublet for large crystals and close triplet for small crystals. The three central components of the 420 ring are not resolved, and the ring then appears as a triplet.

Zinc oxide smoke patterns

The rings from regularly shaped particles other than cubes will similarly have characteristic profiles. In the case of ZnO smoke particles, only the prism faces parallel to the hexagonal axis are developed in the long spines of which this material is composed. Hence the 00*l* rings will be sharp and single, since the streaks due to refraction will lie along the rings; the *hk*0 rings will have streaks perpendicular to the rings, and hence these rings will be broadened; the streaks for *hkl* rings will be inclined at various angles. These points are well illustrated by figure 12, which is an enlargement of part of the inner three rings of a ZnO smoke pattern. From the centre outwards the rings are then 100, 002 and 101, and it will be seen that the inclination of the streaks is as predicted.

Hillier and Baker (1946) have recently published enlargements of ZnO smoke patterns showing similar streaks, which they interpreted as low magnification electron-optical images of the individual ZnO spines. Such images, however, would be radial on the 002 ring and circumferential on the 100 ring, whereas the opposite is the case, so that this interpretation of the streaks cannot be correct.

It is evident from the above that the breadth of the continuous rings in patterns from finer ZnO smoke will be different for different rings; the 00*l*, for instance, will be much sharper than *hk*0 rings.

Irregular particles

Deviation from perfect regularity of, for example, the CdO and MgO cubes will prevent the resolution of the ring components. Selectively broadened rings, in which the components are not resolved (figure 13), were obtained from commercial MgO, shown by the electron microscope to consist of cubes with rounded corners. For completely irregular particles, or the similar case of spherical particles, the refraction effect will broaden all rings by the same amount, as our observations have confirmed. The broadening of the rings for this case may be calculated as follows:—Considering a point reflection due to a single crystal, the intensity at a distance r from the point, in suitable units, will be

$$I_r = \cos\phi_1 \, . \cos\phi_2 \, . \, d\phi_1 \, . \, d\phi_2 .$$

If r_1 and r_2 are the component displacements due to the two faces traversed, $r_1 = q\tan\phi_1$, where $q = P/2E$, and

$$I_r = q^4(r_1^2+q^2)^{-3/2}(r_2^2+q^2)^{-3/2}\,dr_1 \, . \, dr_2 .$$

Taking the radial displacements $p_1 = r_1\cos\gamma_1$ and $p_2 = r_2\cos\gamma_2$, projecting the above distribution on the radius and integrating for all values of γ_1 and γ_2, one obtains for the intensity distribution across the ring

$$I_p = \left(1 - \frac{p_1}{(p_1{}^2 + q^2)^{\frac{1}{2}}}\right)\left(1 - \frac{p_2}{(p_2{}^2 + q^2)^{\frac{1}{2}}}\right),$$

where

$$p = p_1 + p_2.$$

For spherical particles, where $p_1 = p_2$, the half-width of this distribution is $1{\cdot}4\,P/2E$, and the width for one-tenth intensity (approximately the line breadth estimated visually) is $3{\cdot}8\,P/2E$. For values $P = 15$ volts and $E = 50$ kv. these widths correspond to $0{\cdot}6'$ and $1{\cdot}8'$ respectively. For particles of irregular shape $p_1 \neq p_2$ and the half-width will be approximately $1{\cdot}2\,P/2E$.

The above calculation is made for the case of low absorption. For high absorption the central peak of the distribution will be flattened, and the half-width correspondingly greater.

This broadening of the rings due to the refraction effect will be additional to that due to the factors formerly recognized as contributing to ring breadth. These are finite beam width, lack of homogeneity of electron velocities, and finite crystal dimensions. The first two factors may be reduced to negligible proportions by suitable design of apparatus, so that only the last need be considered here. According to Brill (1934) the breadth associated with near-cubic particles of mean cube edge Λ is $0{\cdot}94\lambda \sec\theta/\Lambda$. The broadening due to refraction will be of the same magnitude as that due to finite crystal dimensions for crystal size of the order of 250 A. for $P = 15$ volts and $E = 50$ kv., and the error due to neglect of the refraction effect will still be 10% for 25 A. crystals.

In the case of regularly shaped crystals, the selective broadening of some rings by refraction will be as much as two or three times greater than that for irregular shapes. This will complicate the estimation of crystal shape by comparison of ring breadths. In general, the refraction effect will give differences in the same direction as those due to unequal crystal dimensions. For example, the *hk*0 rings from ZnO smoke will be selectively broadened by refraction and also by the small crystal dimension across the long spines.

§ 6. RELATIVE INTENSITY OF RINGS

For particles of regular shape it will be evident that the peak intensity of a ring which is multiple will be much less than that for an unbroadened ring, although the integrated intensity is identical. In the intermediate case, in which components are not resolved, and for particles of irregular shape, where the ring is simply broadened, a similar situation will obtain. Since ring intensity is usually derived from peak intensities, the intensity distribution will be found to deviate further from those calculated from structure factor for particles tending to assume regular shapes. The unbroadened *h*00 rings for cubes will therefore be relatively strong. Intensity measurements should, of course, always be made by finding the area under ring contours by precise microphotometry. Even so, it may be seen from the ring profiles for 220 and 420 given in figures 11 (*a*) and 11 (*b*) that a considerable proportion of the reflection intensity will, as a result of refraction, appear at relatively large deviations in broadened rings. This part of the intensity will be indistinguishable from background on the photographic plate, with the result that the measured intensities of the broadened rings will be too small by comparison with those of sharp rings.

ZnO smoke gives electron-diffraction patterns which illustrate this effect. Yearian (1935) and others have observed that planes parallel, or nearly so, to the basal plane give reflections that are invariably too strong. Thus reflections from 002, 102 and 103 have intensities greater than theoretical: x-ray diffraction shows the line intensity of 110 > 102 and 103, whereas electron diffraction shows 103 and 102 > 110. This is precisely the way in which refraction for long thin hexagonal prisms would affect the intensities.

It is expected that intensity differences due to refraction will become relatively less important as the crystal size becomes smaller or if resolution is poor. While no quantitative measurements of this effect have been made, the same qualitative intensity distribution has been reported for ZnO smoke examined with cameras of poorer resolving power than that used in this work. Moreover, ZnO specimens prepared by other methods, having less regular particle shape, as shown by electron-microscope examination, give electron-diffraction patterns in which the intensity distribution follows the theoretical closely; in particular, the 110 ring becomes more intense than 102 and 103.

§ 7. FURTHER NEW FEATURES OF ELECTRON DIFFRACTION PATTERNS

As mentioned in § 3 and illustrated in figures 2 and 3, patterns from brown CdO and certain specimens of MgO contain streaks and groups of streaks in place of the expected spots on ring components. The streaks in one group all radiate from a common centre—that is, δ varies in magnitude, but not in direction. The only factor which could account for this type of variation in δ is a variation of the inner potential. Similarly, the second component observed for streaks on the inner rings lies on the same radial line and must therefore be attributed to a second higher value of the inner potential.

In instances where the direction of the streaks represented a recognizable orientation of the crystal so that the crystal faces involved in the refraction could be identified, it was possible to calculate from measurements made in the variation of δ the inner potentials corresponding to various parts of the streaks. In general, the elongation of the inner component of the streaks represented a variation of the inner potential about that found from the arc patterns. For the (220) plane, however, the average value was found to be about 8 volts, somewhat less than the value derived from the 220 arc. The outer component of the streaks corresponds in every case to an inner potential of 25 volts, even for cases in which the potentials corresponding to the inner components differ. The 111 ring of the CdO pattern shows a third component corresponding to an inner potential of 40 to 50 volts.

The variation of the inner potential about the mean value may be interpreted as arising from the presence of excess Cd atoms in inter-lattice sites in the non-stoichiometric oxide. The mean inner potential would change in the region of such defects, and electrons passing through such regions would be refracted accordingly. The presence of components corresponding to inner potentials of 25 volts and more has not yet been explained.

REFERENCES

Bragg, W. H., 1942. *Proc. Roy. Soc.*, A, **179**, 51 and 94.
Brill, R., 1934. *Z. Kristallogr.*, **87**, 275.
Brück, L., 1936. *Ann. Phys., Lpz.*, **26**, 233.
Cochrane, W., 1936. *Proc. Phys. Soc.*, **48**, 723.
Cowley, J. M. and Rees, A. L. G., 1946. *Nature, Lond.*, **158**, 550.
Daniel, V. and Lipson, H., 1943. *Proc. Roy. Soc.*, A, **181**, 368.
Ehrhardt, C. H. and Lark-Horovitz, K., 1940. *Phys. Rev.*, **57**, 603.
Hillier, J. and Baker, R. F., 1945. *Phys. Rev.*, **68**, 98.
Hillier, J. and Baker, R. F., 1946. *J. Appl. Phys.*, **17**, 12.
v. Laue, M., 1936. *Ann. Phys., Lpz.*, **26**, 55.
Lonsdale, K., 1945. *Nature, Lond.*, **155**, 572.
Luneberg, R. K., 1944. *Mathematical Theory of Optics* (Brown Univ.), Appendix II.
Rooksby, H. P., 1943. *Nature, Lond.*, **152**, 304.
Sturkey, L. and Frevel, L. K., 1945. *Phys. Rev.*, **68**, 56.
Yearian, H. J., 1935. *Phys. Rev.*, **48**, 631.

ELECTRON OPTICS AND SPACE CHARGE IN STRIP-CATHODE EMISSION SYSTEMS

By O. KLEMPERER,
Research Laboratories, Electric and Musical Industries, Ltd.,
Hayes, Middlesex

(Now at Imperial College, London)

MS. received 2 September 1946

ABSTRACT. A description of strip cathode systems is given and their merits are compared with those of circular emission systems. It is shown how electron-optical laws can be applied to simple strip systems as long as the emission is sufficiently small. Results obtained under space-charge conditions reveal that the beam spread, due to mutual repulsion of the electrons, completely changes the emission distribution. These results, however, can be interpreted qualitatively by a simple space-charge theory.

Special systems, proposed by Pierce, in which the electron-optical orbits are not upset by space charge, are investigated by measuring the current that can be transmitted through a tunnel of given dimensions. Large discrepancies are found between experimental results and predictions of the simple space-charge theory.

With a view to finding the reasons for these discrepancies, ray-tracing results are discussed and the potential distribution in beams of relatively large space charge is investigated. Tracing results reveal lack of homocentricity, which is an essential supposition in the simple theory of beam spread. The lack of homocentricity, however, is found to be caused by the potential distribution in the beam. This potential distribution is probed in a large beam of high space charge by means of a fine beam of low intensity.

§ 1. INTRODUCTION

STRIP-CATHODE emission systems are electron sources to produce flat, ribbon-shaped or wedge-shaped beams which, by a focusing lens, can be projected into a line focus. The glass-optical counterpart of the strip-cathode system is the elongated light source, the rays of which may be projected into a line focus by means of glass lenses with cylindrical surfaces. The glass-optical and the electron-optical line focus systems can be treated geometrically as two-dimensional problems provided the sources are sufficiently long to render negligible disturbances arising from conditions at the ends. The paths of the rays, however, are entirely different in the two systems, since space-charge effects always play an important part in the electron-emission systems.

Both in light optics and in electron optics the line focus systems have been applied far less frequently than systems of circular symmetry, and for this reason their properties have not received much attention in the past. However, their investigation should be of certain basic interest. Due to the greater simplicity of the two-dimensional geometry, the problems of the strip systems are in many cases more accessible to simple argument. Moreover, a comparison of the different properties of strip systems and circular systems will bring out various interesting points. The voltage-current relations of strip systems and circular systems having the same cross-sectional electrode arrangement are entirely different. For a given cross-section and at a given anode voltage, the strip system will always yield appreciably larger currents, even if the strip cathode is considerably restricted in length. The reason for this is found not only in the relatively larger area of the strip cathode, but also in the relatively larger penetration of fields from the anode to the cathode surface. Also, for a given area of transverse section of the beam and for a given beam current and speed, the disturbance of potential along the beam path due to space charge will be less with a beam of rectangular section than with one of circular section.

The merits of strip systems for the production of large beam currents with relatively small space-charge disturbances have been recognized by Broadway and Bull (1940) who recommended the use of strip beams in high-frequency velocity modulation tubes. In these tubes it is desirable that electrodes for extracting power at high frequency should closely embrace the beam currents, which must themselves be large. Usually a number of such electrodes is necessary, which have to be spaced some distance from each other. All the electrodes have relatively small apertures through which the electron beam has to pass in succession. Since, due to the mutual repulsion of the electrons, it is impossible to produce parallel beams of high current, it is the usual practice to direct an initially convergent beam towards a virtual focus which, however, due to space-charge effects, is never reached by the electrons. The actual beam forms a minimum cross-section or a waist, after which it spreads again. The beam waist has to be located in a certain position with respect to the electrodes of the tube if the passage of a maximum current is desired. The spreading of electron beams due to the influence of space charge and the compensation of it by proper field arrangement in the emission system has been treated by Pierce (1939 and 1940). We shall discuss in the following paragraphs a number of experiments which have been set up with a view to studying the current distribution in strip cathode systems so that the influence of electron optics and of space charge and their mutual interaction can be recognized.

§ 2. SPACE CHARGE AND ELECTRON OPTICS IN SIMPLE STRIP SYSTEMS

The mutual influences of space charge and electron optics upon the electron emission can be studied best in very simple strip systems, in which the two effects can be studied independently and separately as far as possible. One of the simplest available systems of this kind is shown in the side sketch of figure 1. It consists of a plane cathode (Ca) and of two equal slots (Gr) and (An). The sketch represents a cross-section in the (*yz*)-plane in which the beam geometry can be discussed as a two-dimensional problem. All electrodes may be imagined to be extended in the *x*-direction to such an extent that the ends of the electrodes are sufficiently remote from the (*yz*)-plane under consideration for the disturbances caused by end effects to be neglected. If the slot apertures are sufficiently small in comparison with their mutual distance (*a*) and also with the (Gr) slot-to-cathode distance (*c*), the electron-optical effect of each slot upon the beam can be treated separately in a simple manner.

The focal length (*f*) of the line-focus lens formed by a slot is given by Davisson and Calbicks' well-known formula

$$f=\frac{2V}{E'-E}, \qquad \ldots\ldots(1)$$

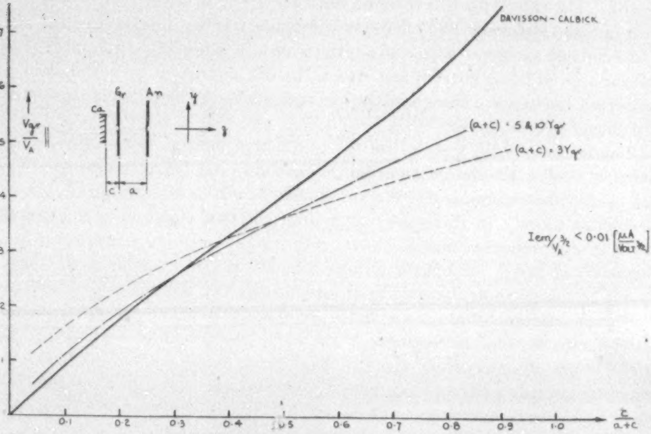


Figure 1. Production of quasi-parallel beams by positive grid bias.

where *E* and *E′* are the electric fields (volts/cm.) in the two spaces which are separated by the slot diaphragm, and *V* is the potential of the diaphragm. C. S. Bull (1945) has investigated recently the application of this formula to aligned grids of thermionic valves, and he has found that the electron paths through the grid apertures were well predicted by the formula. To investigate the application of equation (1) to strip-cathode emission systems, a particular case may be chosen here in order to simplify the experimental tests. According to Bull, the system should emit approximately parallel rays when the gird voltage is positive and reaches a certain fraction of the anode voltage. Calling this fraction

$$n=\frac{V_{gr}}{V_A},$$

formula (1) yields for the grid slot

$$f_1 = \frac{2nV_A}{\frac{V_A(1-n)}{a} - \frac{nV_A}{c}}, \qquad \ldots\ldots(2)$$

a and *c* being grid-to-anode and grid-to-cathode spacings respectively.

For the anode slot

$$f_2 = \frac{2V_A}{-V_A\left(\frac{1-n}{a}\right)}. \qquad \ldots\ldots(3)$$

The condition for parallel rays is

$$f_1 = f_2 + a. \qquad \ldots\ldots(4)$$

From equations (2) to (4) it follows that

$$n = \frac{\left(6+3\frac{a}{c}\right) \pm \sqrt{9\left(\frac{a}{c}\right)^2 + 24\frac{a}{c}}}{6+2\frac{a}{c}}. \qquad \ldots\ldots(5)$$

The voltage ratio *n* as a function of the ratio of spacings $c/(a+c)$ which produces parallel rays according to equation (5) is plotted as a solid curve in figure 1. Only the negative root is considered, since, for practical purposes, we are interested only in values of *n* less than 1. As a purely electron-optical result derived from equation (5), the first slot should always have positive potential for producing parallel rays. On the other hand, experimental evidence shows that under ordinary conditions, i.e. under space charge, a two-slot system always produces divergent beams, the angle of divergence being a minimum when the grid is at zero or at negative bias. It follows, therefore, that under normal conditions of space charge, the mutual repulsion of the electrons is responsible for the divergence of the beam.

The magnitude of space-charge effects in an electron beam of the current (I) with the energy (V) is controlled by the "space-charge factor" ($I/V^{3/2}$). We derive this from Child's law, according to which, in emission systems of any geometry, the electron current grows about proportionally with the 3/2 power of the anode voltage as long as full space-charge conditions are maintained. If we want to get rid of space-charge effects, we have to reduce the emission current (I_{em}) or to increase the anode voltage (V_A) of our system in such a way that $I_{em}/V_A{}^{3/2}$ is reduced. A reduction of $I_{em}/V_A{}^{3/2}$ can be obtained by underheating the filament of the cathode, i.e. by decreasing the cathode temperature.

If in this way the space-charge factor of the simple two-slot emission system was reduced to less than 1/100 of its full value, it was found experimentally that the ordinary electron-optical laws expressed by equations (1) and (5) could be applied. In the experimental test of equation (5) the ratio of grid to anode voltage

$$n = \frac{V_{gr}}{V_a}\Bigg\|$$

for producing parallel rays had to be found as a function of the spacings *a* and *c*. A fairly good estimate of the angle of divergence, or, in particular, of the parallelism of the electron beam, could be obtained with a fluorescent target at a sufficiently large distance from the anode slot of the emission system.

In the actual measurements, the bias at the grid diaphragm was varied until the least beam divergence could be observed. This was taken as an indication of the production of approximately parallel beams. The ratio of this particular grid bias to the anode voltage ($=V_{gr}/V_A\|$) is plotted in figure 1 against the ratio of cathode to grid spacing (c) over cathode-to-anode spacing ($a+c$). The broken curve applies to an ($a+c$) spacing equal to three grid-slot-semi-apertures y_{gr}. The full curve nearest to the dotted one represents the theoretical values of equation (5). Complete agreement with the experimental values could not be expected, since the experimental grid electrode was of finite thickness (1/3 of the width of the grid-slot-aperture); this thickness, however, is neglected in Davisson and Calbick's formula. However, there seems to be no doubt that under the conditions of figure 1, the paths of the electrons are controlled entirely by electron-optical laws.

The measurements refer to the electrode potentials to produce quasi-parallel beams. These beams were not sharply defined but the intensity was strongest in the middle and faded away to the edges. The intensity distribution in the beam is produced by the velocity distribution of the emission and the field distribution at the cathode surface. The velocity distribution is known to be Maxwellian; the field distribution can be calculated or measured in the electrolytic trough. The actual measurement of this distribution and its comparison with the theory would be laborious and of little interest.

Practical emission systems are very rarely used under conditions of extremely low space charges. With increasing space-charge factors, however, the minimum obtainable beam spread increases rapidly. The grid bias, at which the minimum spread is obtained, changes with increasing space charge from positive to zero and to negative values. The intensity of the beam is strongest in the middle, and towards larger angles it appears to fade away so gradually that the edges are not clearly defined.

The current distribution of a beam originates at the cathode, and the emission distribution there may reach a certain equilibrium in its interaction with a particular space-charge distribution. The current distribution of an intense beam is, at a later stage, still decisively influenced by the mutual repulsion of the electrons on their paths. The most essential results on the beam spread can certainly be derived from this mutual repulsion of the electrons. We shall proceed, therefore, to give a short outline of the simple space-charge theory of the beam spread, and we will see how far it can be applied to an interpretation of the experimental results obtained with simple strip systems.

§ 3. APPLICATION OF SIMPLE SPACE-CHARGE LAWS TO STRIP SYSTEMS

A simple theory of the spread of ribbon-shaped beams has been developed by A. Bouwers (1935) and by J. Thompson and L. B. Headrick (1940). This theory can be applied to all parts of the beam in which the forward velocity of the electrons is constant, and in which the direction of the electrons is strictly homocentric in the yz-plane, in which the problem can be treated as a two-dimensional one. There, each electron path describes a parabola, the parameter of which depends upon the initial convergence and upon the space-charge factor ($I/V^{3/2}$) of a beam, the boundary of which is defined by the parabola considered. According to Bouwers, we can distinguish the three main types of paths which are shown in figures 2, 3 and 4 of the present paper.

In all three figures, the beam (El) enters at the co-ordinate (z_A) through a slot of the semi-aperture (y_A). At the slot the beam is directed at an angle θ with the z-axis towards a virtual focus at z_F, which, however, is never reached because of the mutual repulsion. In figure 2 the beam reaches a waist, i.e. a minimum semi-aperture y_W at z_W, while the beam in figure 4 forms a real cross-over with zero aperture at z_X. Figure 3 shows the border case or transition between the types of figure 2 and figure 4. The point of minimum cross-section ($y_W = 0$) at ($z_W = z_X$) can be considered to be the transition between a waist and a cross-over. Its distance from the aperture is here a maximum. The equation of the electron path is

$$y = y_A - \theta z + \frac{1}{4k} I/V^{3/2} z^2. \qquad \ldots\ldots(6)$$

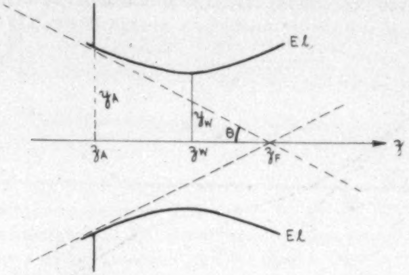


Figure 2. Electron path and space charge.

The aperture being taken as the origin, i.e. $z_A = 0$, from equation (6) follows

$$\left.\begin{aligned} y_W &= y_A - \frac{\theta^2 k}{I/V^{3/2}}, \\ z_W &= \frac{2k\theta}{I/V^{3/2}}, \end{aligned}\right\} \qquad \ldots\ldots(7)$$

and with $y_X = 0$

$$z_X = \frac{\theta \pm \sqrt{\theta^2 - \dfrac{y_A I}{kV^{3/2}}}}{(I/V^{3/2})/2k}, \qquad \ldots\ldots(8)$$

where k is a constant. The numerical value of this constant is 10·4 if the current (I) is measured in microamp. per cm. length in the x-direction, i.e. for a width of the strip-beam $2x_A = 1$ cm. Putting $y_W = 0$ in equation (7) gives for the critical case of figure 3 the following critical space-charge factor:

$$I/V^{3/2} = \frac{k\theta^2}{y_A}. \qquad \ldots\ldots(9)$$

Moreover, since $\theta = y_A/z_F$, it follows from equation (7) for any beam convergence

(θ) and any semi-aperture (y_A) satisfying the critical condition equation (9), that the waist distance from the aperture must be just twice as large as the distance of the virtual focus, i.e. $z_W = 2z_F$.

Figures 2 to 4 are all drawn to scale in the y co-ordinate, but scaled down by a factor 10 in the z co-ordinate in order to make the essential features more visible. The space-charge factor in all these curves is taken as $I/V^{3/2} = 0{\cdot}1$ (microamp./volts$^{3/2}$), but the angle (θ) is chosen as 0·05, 0·1 and 0·12 radians respectively in the three figures. The values of y_W, z_W and z_X, plotted in the figures, are calculated according to equations (7) and (8), the semi-aperture (y_A) being taken as unit length. While figures 2 and 3 need no further explanation, it may be pointed out that the broken curve in figure 4 represents the parabola given by equation (6); this is continued in the drawing after the z-axis is reached at z_X. The real electron path, however, is shown as the solid line (El); it is the reversed first branch of the parabola between (y_A, z_A) and (0, z_X), plotted in ($-y$) direction. It may be emphasized here that the rays in strip-beams of sufficiently large angle of convergence (θ) and sufficiently small space-charge factor ($I/V^{3/2}$), as shown in figure 4, really cross over. In distinction to this, a real cross-over cannot be obtained in beams of circular cross-section, unless the beam currents are so much decreased that discontinuities of space charge due to its composition of single electrons start to play a part.

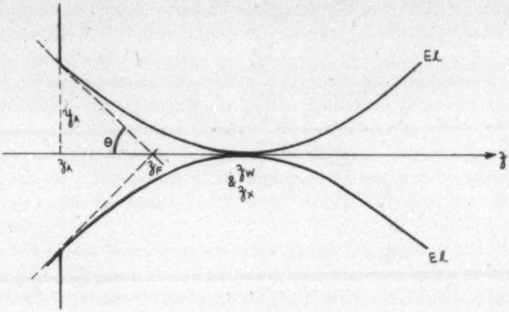


Figure 3.

In the attempt to compare Bouwers' simple theory and the experimental results obtained with simple strip-cathode systems, we are obviously not interested in details of the angular distribution of the emission caused by the field distribution over the cathode surface. On the other hand, if some function of the space-charge factor could be found to represent the beam spread, at least a qualitative comparison would be possible between theoretical and experimental results. In a somewhat arbitrary way we have chosen two characteristic measures for the beam spread which apply to parts of the beam sufficiently far away from the waist. There, the beam has spread sufficiently, so that space charge has hardly any further effect upon the paths of the electrons, which approach straight lines. We have measured at this point either the semi-vertical angle (θ_h), which includes half the beam current, or we have measured the fraction of the total emission I_{ya}/I_{em}, which is projected into a semi-vertical angle of 0·1 radians.

For both types of measurement an adjustable slot (Sl) in front of a Faraday cage, as shown in figure 5, was used. There, (Ca), (Gr) and (An) represent the strip-cathode, the slotted grid electrode and the anode respectively.

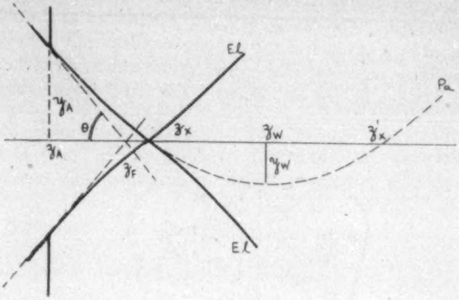


Figure 4.

The slot (Sl) was adjustable in the vacuum; its design was developed from an optical model described by Sears (1933) and by Strong (1941). A schematic diagram is given in figure 6. The brass frame (Fr) was mounted on the four-rod assembly (Rd) on to which the whole gun was fixed. On this frame were pivoted two pairs of parallelogram arms (Pa) holding the jaws (Jw) on to which were screwed the two thin slot-blades (Bl). The jaws were pushed forward against the spring-blades (Sp) by a sliding part (Bp) which was guided in the two blocks (Ga) and (Gb). In order to vary the slot width, the sliding part was screwed forward by the internally threaded helical gear-wheel (Ha) which was rotated and driven by

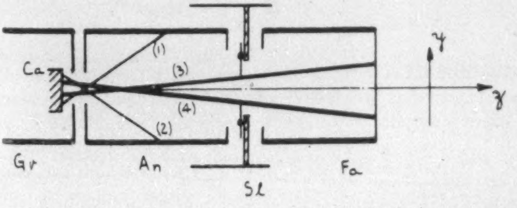


Figure 5. Measurement of angular current distribution.

another helical gear-wheel (Hb), arranged at right angles to it. The wheel (Hb) was fixed to an axle which was connected to a ground joint so that it could be turned from outside the tube. The resulting movement of the jaws covered changes of semi-aperture up to $y = \pm 4$ mm. The zero-slot width was adjustable by the position of the blades which were screwed on the jaws.

Experimental curves of the current efficiency, i.e. (I_{Fa}), as a function of (θ) for small θ did not deviate much from a straight line (up to $\theta = 0{\cdot}1$). Only for large θ these curves gradually bend round at some point to converge towards the 100% efficiency, when all available electrons are included inside the considered wedge.

Figure 7 shows the change of beam spread caused in a given system by the change of the space-charge factor only, the grid electrode being constantly kept at cathode potential. The change of $I_{\mathrm{em}}/V^{3/2}$ is plotted as abscissa, this change being produced by a variation of the cathode temperature. The beam spread is characterized by the angular semi-aperture (θ_{h}) into which half of the total emission is projected. (θ_{h}) is plotted as ordinate. For small space-charge factors, the electrons follow, as we have seen, purely electron optical laws. The equipotentials near the grid electrode are strongly curved, thus the electron paths cross over steeply and the resulting beam is very divergent. With increasing space charge, the repulsion of the electrons at the cathode side of the cross-over tends to reduce the initial convergence of the beam and thus reduces its final divergence. Thus the left-hand part of the curves of figure 7 is explained by the transition of the beam from the stage characterized by figure 4 to the stage shown by figure 3. The curves of figure 7 pass through a minimum and rise again when, by increased space-charge effects, the initial convergence is further reduced with a simultaneous increase of the divergence after the beam has passed the waist. These conditions correspond to the stage represented by figure 2.

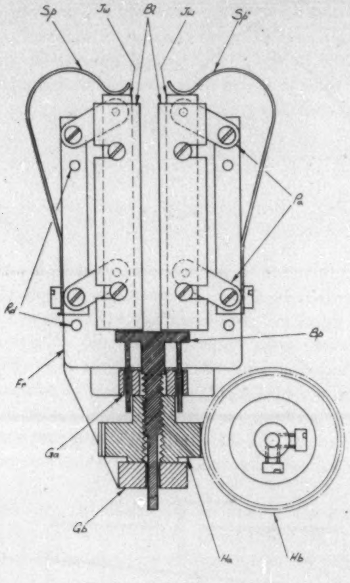


Figure 6. Parallelogram slot.

Bouwers' simple theory, as illustrated by figures 2 to 4, is adequate for a rough qualitative explanation of the essential facts. To understand details about beam spread and angular distribution, the following considerations have to be given:—

1. In every cathode system the space-charge factor $(I/V^{3/2})$ of the beam decreases gradually on the way from the cathode, where (V) is very small, to the anode, where $V = V_A$. Bouwers' theory, however, applies to beams with homogeneous space-charge factors.

2. It has already been pointed out in § 2 that the current distribution of the beam originates at the cathode surface, which is exposed to a certain distribution of field strength. For intense beams, however, this field distribution is produced not only by the electron-optical potential distribution but also by the space-charge distribution in front of the cathode. Eventually, the emission distribution reaches a certain equilibrium in its mutual interaction with the space-charge distribution.

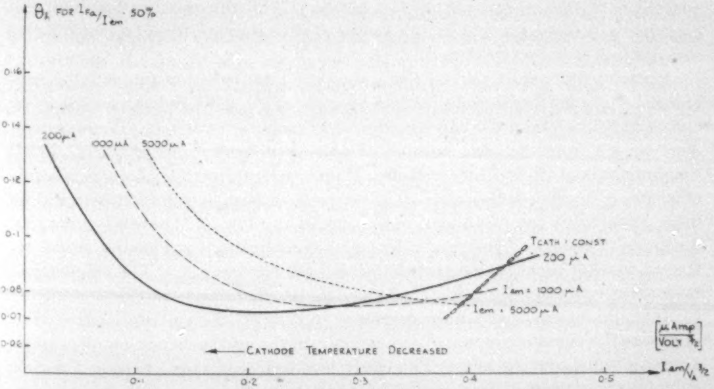


Figure 7. Beam spread and space charge open anode system.

3. The initial current distribution of the electrons at the cathode is modified by an initial thermal-velocity distribution and by the influence of space charge upon this distribution (*cf.* Klemperer, 1947). The influences arising from the velocity distribution, however, are of secondary importance for the current distribution and they will not be studied in the present paper.

4. The potential distribution set up in the beam by the space charge produces a decrease of electron velocities in the paraxial parts of the beam. This reacts again upon the potential distribution until equilibrium is reached. We shall deal with this effect in § 5 of this paper.

Changes in angular current distribution produced by changes in field distribution at the cathode surface as mentioned under (2) above are very noticeable in the transition region between temperature saturation and space-charge condition as in figure 7. There, the current distribution of the beam not only changes with a change in ($I_{em}/V_A^{3/2}$), but even if this latter space-charge factor is kept constant, it changes with the anode voltage (V_A). There is on the cathode surface a central zone with temperature-limited emission. The width of this zone apparently depends upon the absolute value of the field. As a consequence, the three curves of figure 7 which were measured at 0·2, 1 and 5 milliamp. differ appreciably. The change of field distribution at the cathode surface is also responsible for the fact that at a constant cathode temperature, and even under conditions of full space charge, the factor ($I_{em}/V_A^{3/2}$) decreases appreciably with increasing anode voltage or emission current. A short curve, which in figure 7 is indicated by a chain of small circles, connects the points of equal cathode temperature under full space-charge conditions in the curve of beam spread against space-charge factor.

The curves of figure 7 are instructive in showing the powerful influence of the gradually increased space charge. In practical applications, however, emission systems are always used under full space-charge conditions. There, both the full space-charge factor at the anode, which is usually called the *perveance* of the system, and the divergence of the emitted beam can largely be controlled by the spacing between cathode and grid electrode (c), by the spacing between grid electrode and anode (a), and by the widths of the semi-apertures (y_{gr}) of the grid slot and (y_A) of the anode slot.

Experimental values for the fraction of the total emission under full space-charge conditions that is projected into an angle of $\theta = \pm 0{\cdot}1$ radians are shown by curves in figure 8 for a few characteristic representative systems. These systems were chosen from the great number of possible combinations of plane, slotted diaphragms and of a plane cathode. This current fraction characterizes the efficiency of the emission system. It is, for instance, high (0·8) for a system with an open, wide anode (no diaphragm in the anode) and low (0·24) for a symmetrical system with equal grid and anode slot (semi-apertures $y_A = y_{gr}$) having an anode-to-grid-spacing equal to one grid slot semi-aperture ($a = y_{gr}$). The efficiency of both these systems is not markedly dependent on the cathode-to-grid spacing (c).

It would be expected that for decreased spacing (c), the electron-optical effect of the smaller curvature of the equipotentials near the grid electrode would produce a smaller initial convergence. This effect, however, seems to be balanced to some extent by an increase in beam spread due to an increasing space charge, the space-charge factor being larger the smaller the spacing (c). If, now, with further decrease of (c), the space-charge factor is further increased, beam conditions will eventually reach the critical stage represented by equation (9). We can then expect a minimum in beam spread, i.e. a maximum in efficiency. The critical spacings of the electrodes, at which this maximum occurs, have been measured and the particular conditions have been discussed sufficiently in a paper on emission systems with circular symmetry by Klemperer and Mayo (1947), so that we need not enlarge further on this subject. We shall only point out here some characteristic differences between the circular and the strip systems.

It has been pointed out elsewhere (*cf.* Klemperer, 1939, p. 96) that two-dimensional (line focus) lenses always have shorter focal lengths than three-dimensional circular-symmetry lenses of the corresponding electrode arrangement. Thus the critical conditions which for symmetrical circular two-diaphragm systems were found to occur at a critical grid-to-cathode spacing of the order of a grid semi-aperture ($c \approx y_{\mathrm{gr}}$) would be expected for the corresponding slot system at much smaller spacings. This is borne out by the two dotted curves of figure 8, which do not show maxima within the plotted range. Apparently, for technical reasons, the spacings (c) from the top of the grid to the cathode could not be made small enough to reach the critical conditions.

On the other hand, an unsymmetrical two-slot system, with an anode slot half as wide as the grid slot ($y_{\mathrm{A}} = y_{\mathrm{gr}}/2$), would have electron-optically a greater focal length and, in addition, would have greater perveance ($I_{\mathrm{em}}/V_{\mathrm{A}}^{3/2}$) than the corresponding symmetrical two-slot system. Now, grid-to-cathode spacings leading to the critical conditions equation (9) can be easily realized. The broken curves in figure 8 belong to such an unsymmetrical two-slot system and they show distinctly the expected maxima. The three curves shown were all taken under full space-charge conditions, but at the three different emission currents of 1, 5 and 20 milliamp. The wide difference between these three curves points to a relatively large difference in field distribution over the cathode surface, which seems to have a particularly great effect upon the angular current distribution under the critical conditions near the maximum efficiency. In this respect, slot systems behave analogously to the circular systems; however, in contrast to the circular systems, even near the critical stage, the slot systems are far less subject to the influence of residual gas.

Concerning the systems with open box anode the critical conditions of equation (9) may be obtained by reducing appreciably the dimensions of the anode. This can be seen by comparison of the two full curves of figure 8; no further comments, however, need be given about this case.

As a point of practical interest, it may be mentioned that the perveance of the different types of slot systems, even for a given length of slot, may be of a very different order. The perveance of course changes very rapidly with the spacings a and c, but even if we fix both these spacings at $1y_{\mathrm{gr}}$, we obtain the following rather different values: $I_{\mathrm{em}}/V_{\mathrm{A}}^{3/2} = 0\cdot1$ μamp./(volt)$^{3/2}$ per cm. slot length for the slot system with wide, open anode, but $5\,\mu$amp./(volt)$^{3/2}$ per cm. slot length for the unsymmetrical two-slot system. In comparison, the circular symmetrical two-diaphragm systems and open anode systems reach only $I_{\mathrm{em}}/V^{3/2} = 0\cdot2$ and $0\cdot006$ μamp./volt respectively.

§ 4 EMISSION SYSTEMS OF PIERCE'S TYPE

The discussions of §§ 2 and 3 apply to the simplest type of strip-cathode emission systems consisting of plane electrodes only. It was shown how, in these systems, space-charge effects completely upset the particular orbits of the electron rays that follow from purely electron-optical laws, and that are realized only at vanishingly small space-charge factors. The field distribution at the cathode surface further complicates the angular current distribution in the beams emitted from the above simple systems.

In order to have theoretically simple conditions which would allow the highest possible beam concentrations, Pierce (1939 and 1940) proposed a new kind of emission system in which the electric field is constant over the whole cathode surface and in which space-charge effects would not tend to alter the electron-optical paths of the electrons. In Pierce's strip system, cathode and grid electrode are parts of two coaxial cylinders. It is known that between two complete coaxial cylinders a strictly radial flow of electrons is obtained whatever conditions of space charge are given. If a sector is taken out of the complete cylinder, the radial flow of the electrons can be preserved if the potential distribution along the beam is kept unchanged everywhere, especially at the boundary face, which now occurs between electron beam and empty space. Pierce suggested producing at this boundary—by means of suitably shaped electrodes—a field which agrees with the space-charge conditions, i.e. which shows along this boundary a potential distribution proportional to the 4/3 power of the distance from the cathode.

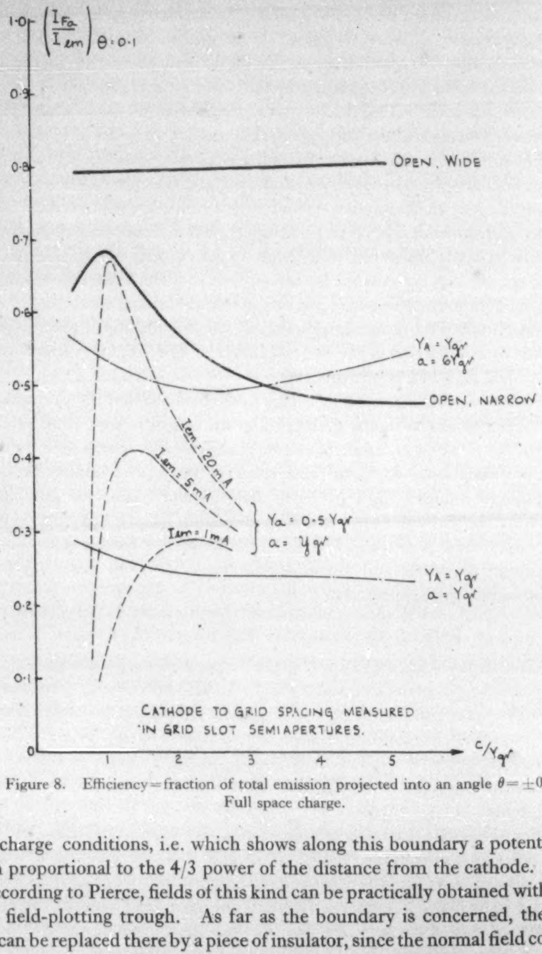


Figure 8. Efficiency = fraction of total emission projected into an angle $\theta = \pm 0{\cdot}1$. Full space charge.

According to Pierce, fields of this kind can be practically obtained with the help of the field-plotting trough. As far as the boundary is concerned, the electron beam can be replaced there by a piece of insulator, since the normal field component in the electrolyte at the boundary of the insulator will vanish just as the normal field component in the vacuum vanishes at the edge of the electron beam. Figures 9 and 10 show two different cathode systems, both having been shaped according to experiments with large-scale models in the field-plotting trough so as to satisfy Pierce's conditions.

The system shown in figure 9 has a concave cathode (Ca) which is surrounded by the cathode shield (Sc), both sides of which are fixed to the two field-shaping shields (Sg). Opposite to (Ca) there is arranged a wire grid anode (Wa) of circular cross-section in the *yz*-plane, which bears the two field-shaping shields (Sa). Electrons, all homocentric in the *yz*-plane, are supposed to pass through the wire grid (Wa) and to enter the field-free tunnel (Dr) through the slot (B); they leave (Dr) through a slot (E) equal in size to (B).

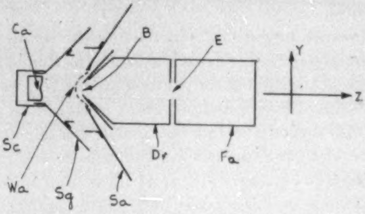


Figure 9. "Pierce gun" and "electron tunnel."

We should expect to get a maximum current through a tunnel formed by two equal slots if the parabolic electron orbits were arranged symmetrically to it, i.e. if the beam forms its waist exactly in the middle of the tunnel. Moreover, space-charge factor and angle of convergence (θ) at the aperture should be chosen so as to produce the critical case represented by equation (9) or figure 3. Now θ is fixed by the geometry of the tunnel, i.e. by the width of the slots ($2y_A$) and their mutual distance ($2l$). Since, as we explained above for the critical case, the

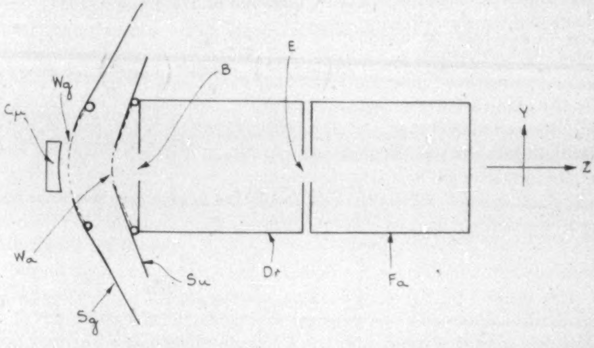


Figure 10. "Pierce gun" and "electron tunnel."

distance of the virtual focus from the aperture (z_F) is just one-half of the waist distance ($1/2\,z_W$) from this aperture, we have

$$\theta = \frac{y_A}{z_F} = \frac{2y_A}{z_W} = \frac{2y_A}{l} \qquad \ldots\ldots(10)$$

and with equation (9)

$$I/V^{3/2} = \frac{8ky_A x_A}{l^2}, \qquad \ldots\ldots(11)$$

where $2x_A$ is the breadth of the slot, which is not written down expressly in the notation of equation (9), since this equation applies to the unit beam breadth.

If we require the total output of the cathode system to pass through the tunnel, the system has to satisfy the requirements given in equations (10) and (11). Bull and Klemperer (1944) drew the further conclusion that in cathode systems such as shown in figures 9 and 10, the curvature of the cathode and the cathode-to-anode distance are both fixed by equations (10) and (11) as soon as the tunnel dimensions are given. This can be seen immediately from Child's equation which—according to B. J. Thompson (1943)—can be written for a cylindrical cathode system, expressing again the total emission I_{em} in microamp. and the anode voltage V_A in volts:

$$\frac{I_{em}}{V_A^{3/2}} = 2{\cdot}33\,\frac{A}{r_A^2\beta^2} \qquad \ldots\ldots(12)$$

where A, the area of the cylindrical anode, is given accurately enough by $(2y_A)$ as long as the arc can be replaced by the chord $(2y_A)$. Moreover, r_A represents the radius of curvature of the anode which is here $r_A = z_F$. β is a function of (r_A/r_c), the ratio of the radii of curvature of anode and cathode. The function β is given in a graphical representation by I. Langmuir and K. T. Compton (1931), where it is approximated by the following series:

$$\beta = \log_e \frac{r_A}{r_c} - 0{\cdot}40\left(\log_e \frac{r_A}{r_c}\right)^2 + 0{\cdot}092\left(\log \frac{r_A}{r_c}\right)^3 . \qquad \ldots\ldots(13)$$

On the other hand, β is given by substitution of $I/V^{3/2}$ of equation (11) for $I_{em}/V_A^{3/2}$ into equation (12) since the cathode system is supposed to produce a beam of exactly the space-charge factor which the tunnel requires. Thus

$$\beta^2 = \frac{2{\cdot}33}{z^2_F}\,\frac{(2y_A x_A)\,(2z_F)^2}{(4y_A x_A)\,k} = 0{\cdot}448$$

which, with equation (13), yields

$$\frac{r_c}{r_A} = 1{\cdot}71 = \frac{z_F + d}{z_F} \quad \text{or} \quad \frac{d}{z_F} = 0{\cdot}71,$$

i.e.

$$d = 0{\cdot}35l, \qquad \ldots\ldots(14)$$

where $d = r_c - r_A$ is the distance between cathode and anode.

It follows that we should expect to pass the total emission of the cathode through a tunnel of relatively small aperture as soon as (1) the anode radius is a quarter of the length of the tunnel, (2) the cathode-to-anode spacing is 0·17 times the length of the tunnel.

The system shown in figure 9 was assembled in exactly these required geometrical proportions. As a further precaution, the strip cathode (60 × 6 mm., radius 10 mm.) was covered up at the ends to cut out the regions of disturbance according to a scheme tried out in earlier experiments. $I_{em}/V_A^{3/2}$ was as expected, and its measured value was about 25. Less than 20 % of the cathode emission, however, was received in the Faraday cage (Fa). Only a few per cent of this emission was caught at the wire grid (Wa), which consisted of 60 parallel 0·1 mm. dia. molybdenum wires welded across the anode slot and at the jaws of the entrance slot (B) of the tunnel. Most of the emission current was apparently caught at the diaphragm containing the exit slot (E) of the drift space. Improved yield (35%) could be obtained by changing the cathode-to-anode distance to an optimum of 6 mm., reducing thus the perveance $I_{em}/V_A^{3/2}$ to 12 microamp./volt$^{3/2}$.

There was some reason to suspect disturbances of the field caused by the unavoidable gap between the grid shield (Sg) and the cathode. These disturbances were avoided by another construction shown in figure 10. All inscriptions at this figure correspond to those in figure 9. However, a grid (Wg) (again consisting of 60 molybedenum wires of 0·1 mm. dia. welded in 1 mm. mutual distance across the slot) was inserted closely in front of the cathode.

The grids were pressed to the correct radii of curvature, with the help of specially made jigs, in order to obtain the necessary accuracy in agreement with the theoretical requirements. The single wires of the grids (Wg) and (Wa) were aligned, and in order to avoid undesirable electron focusing by Wg, this grid was kept at a slightly positive bias, which was found empirically by adjusting it to obtain the best yield in the Faraday cage. The current efficiency of this system (figure 10) was not better than that of the system of figure 9.

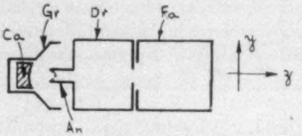


Figure 11. Empirical cathode system for electron tunnel.

When we varied the radii of curvature of the grids Wg and Wa, we noticed, however, that the efficiency could be improved to 60 % of the total emission when the grids were not coaxial, but the radius of curvature of Wg was made smaller and the radius of curvature of Wa larger than their respective distances from the required virtual focus, z_F.

In the end, by a purely empirical development, we arrived at the construction shown in figure 11. There, concave cathode (Ca) and grid shield (Gr) were similar to the corresponding electrodes in figure 9. The entrace of the tunnel (Dr) was formed by the anode nozzle (An). The cathode system had a virtual focus near the end of this anode nozzle. Inside the nozzle, however, there was a strongly diverging field produced by concave equipotentials penetrating into it. Due to the simultaneous action of this diverging field and of the mutual repulsion inside the beam, a waist was probably formed in the middle of the drift space. The optimum anode-to-cathode distance was found empirically by shifting the cathode and grid electrode (Ca and Gr) by a micrometer bellows gear and observing the current in the Faraday cage (Fa). This current could also be adjusted to an optimum by varying a small voltage applied to the grid electrode. This bias voltage was found to equal cathode potential for the above optimum and was found to be negative for shorter distances and positive for larger distances. Apparently the bias controlled the waist distance from the cathode; probably the waist was formed in the middle of the tunnel for the correct combination of cathode position and bias. Moreover, the application of a fine wire grid inside the anode nozzle has been tried. It was found that no disturbances were caused if the grid was strongly concave towards the cathode, as is indicated by the dotted line in figure 11. If, on the other hand, a convex grid was used, such as shown in the guns of figure 9 and figure 10, the current efficiency was greatly reduced.

From experimental results of this kind, it can be concluded that in order to transmit a beam of the greatest possible space-charge factor through a narrow tunnel of given dimensions, this beam should be caused to be "initially over-focused", i.e. it should aim at a virtual cross-over appreciably nearer to the cathode than would be required by equation (10). Then, entering the tunnel, the beam should be spread out by passing through a series of equipotentials which are concave facing the cathode.

In this way particularly compact emission systems can be designed in which the distance between cathode surface and beam waist is relatively short. However, the shorter this distance can be made the less pronounced will probably be the longitudinal potential distribution in the beam and the beam spread produced by it. The potential distribution in the beam will be discussed in detail in the next section of this paper.

The system of figure 11, which represents an optimum design, had a current efficiency of 80%. This implied that a current of the space-charge factor of 7 microamp./(volt)$^{3/2}$ passed through a tunnel of the dimensions $x_A = 30$, $y_A = 1{\cdot}5$, $l = 10$ mm. This, however, represented only 18% of the space-charge factor (38 microamp./volt$^{3/2}$) which was expected from theoretical reasons to pass through this tunnel according to equation (11). The "Pierce" system of figure 9 passed only 10% of the theoretical value through its tunnel. Moreover, the best empirical system that could be made up from a plane strip-cathode and two plane, slotted diaphragms, such as described in § 2 of this paper (*cf.* maximum of broken curve in figure 8) had an efficiency of 45% through a tunnel of similar proportions, but the maximum transmitted $I/V^{3/2}$ amounted only to 17% of the theoretical optimum expected from equation (11).

The current efficiency is to some extent critical with respect to alignment and adjustment, and the chance that a still more efficient type of system might be found by future research cannot be excluded. However, the great number of results obtained with very different types of systems suggests that even under the most favourable conditions a given tunnel could in practice not pass more than, say, a quarter of the space charge which is calculated by equation (11) on the simple theory outlined in § 3. This statement refers to slot systems. Analogous experiments with circular systems lead to similar conclusions. However, the upper limit of the ratio of practical to theoretical space-charge factor of the current passed through a circular tunnel was decidedly better and was estimated to be about 1/2 for tunnels in which diameter and length were of the same order of magnitude.

§5 REASONS FOR THE INADEQUACY OF THE SIMPLE THEORY OF BEAM SPREAD

In the attempt to find reasons for the quantitative discrepancies between the expectations of the simple beam spread theory, outlined in §3, and the results given in §4, certain indications have been obtained from electronic ray-tracing experiments. A divergent beam emitted from a cathode system may be intercepted by a pepperpot diaphragm and the pencils may be traced with microscope and sliding fluorescent target (*cf.* Klemperer and Wright, 1939; also Klemperer and Mayo, 1946). If the cross-over of such a beam is extrapolated, it appears that the rays do not emerge from a common centre. Moreover, if the cathode temperature is changed, the pencils generally become increasingly less homocentric with increasing space-charge factor. Space charge appears to introduce positive spherical aberration. This is explained in figure 12. There, two marginal rays (Ma) and two paraxial rays (Pa) emerging through a pepperpot diaphragm (Pp) are extrapolated backwards. Their virtual cross-overs do not coincide, i.e. the bundles are not homocentric. From these observations, the actual waists (Wm) and (Wp) of the marginal and paraxial rays respectively can be located approximately. Apparently, the marginal waist is formed first and the paraxial waist later.

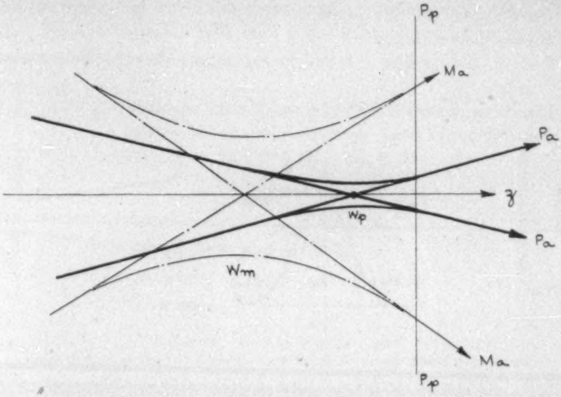


Figure 12. Paths of electrons in beams of high current density.

The diagram in figure 13 represents a potential distribution across the beam. Due to space charge, the negative potential difference $(V_A - V)$ rises beyond the anode voltage (V_A) towards the middle of the beam. This is indicated by the broken line, which could be calculated for a homocentric bundle under the assumption of constant space-charge density over the *xy* cross-section. The potential V_0 at the axis can be derived from Gauss's law and is given by the following equation:

$$V_A - V_0 = \frac{\pi y_A}{\epsilon_0 \sqrt{\frac{2e}{m}}} \frac{I}{V_A^{1/2}} = 4\cdot76 \times 10^{-2} y_A \, V_A \frac{I}{V_A^{3/2}}, \quad \ldots\ldots(15)$$

where the beam current I is measured again in microamp. per cm. breadth, the potentials V in volts and the beam semi-aperture y_A in cm. ϵ_0 is the dielectric constant. It can be seen that the potential hump in the middle of the homocentric beam grows proportionally with the semi-aperture of the beam and with the beam current, but inversely proportional with the square root of the beam voltage.

Due to this rise in negative voltage, however, the paraxial part of the beam is slowed down, i.e. its space charge is increased with respect to the space charge of the marginal beam. As a consequence, the paraxial potential will be further raised, and so on, until a new equilibrium is obtained which is indicated in figure 13 by the solid curve. It can be seen from the figure that, in the new curve, the paraxial potential gradient has risen more than the marginal one. Moreover, due to the increased paraxial space-charge factor, the position (z_W) of the paraxial waist is shifted according to equation (7). As a consequence of the paraxial voltage change there would be thus expected:

(1) Loss of homocentricity.
(2) Loss of homogeneity of beam current density through the cross-sectional area.
(3) Loss of symmetry of the beam waist in the yz plane.
(4) Dependence of beam geometry upon anode voltage.

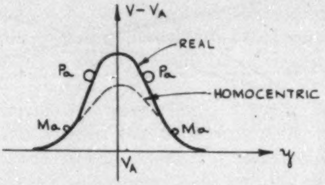


Figure 13. Lateral potential distribution in beam of high current density.

Even if the space-charge factor were kept constant, the paths of the electrons could no longer be considered to be invariant with respect to voltage changes. The paraxial voltage hump would grow according to equation (15), proportionally with the beam voltage, if $I/V_A^{3/2} = \text{const.}$

A theoretical investigation of the potential distribution in an electron beam due to its space charge has been published by Smith and Hartman (1940), and some approximate expressions have been presented for the spreading of circular beams. According to these authors, the velocity distribution of the electrons, due to the space charge, causes the beam to spread more rapidly than can be concluded from the simple theory (*cf.* Watson, 1927). The circular beam will diverge a given amount in about one-half of the distance computed from the simple equations of Watson. Our experimental results, however, neither lead directly to the beam spread curve nor to the potential distribution.

Since a knowledge of the potential distribution in the beam appears to be of basic importance for the understanding of the actual space-charge effects, we have started to investigate it by direct measurements. As a first step, we have obtained qualitative results for a potential distribution across and along a spreading ribbon-shaped beam.

Wehnelt and Bley (1926) first showed that the space charge in front of a cathode can be probed by a fine electron beam. We have adapted their procedure for our purposes and our experimental arrangement is shown in figure 14. There, a very fine strip beam (El) emerging from the anode (An) of an electron gun was used to probe the potential distribution of a space charge emitted from the cathode (Ca). The deflection of the electrons of this probe beam by the space charge was eventually measured on a fluorescent target (Ta). The position of the probe beam was adjusted by the deflection plates (Df). The probe beam while passing the space charge was only about 0·1 mm. wide. A beam velocity of 500 volts was found to be convenient. The probe beam passed through the open sides of an anode-box formed by the anode slot (As) and the anode roof (Ar) of a "space-charge producing" gun. The latter contained a two-slot emission system such as has been described in §§ 2 and 3 of this paper. This system had a plane cathode (Ca), a flat grid electrode (Gr) with a slot aperture and a slotted anode (As). The grid electrode (Gr) was shielded by a wire gauze (Sh) connected to the potential (V_A) of the anode (As) and (Ar) to which was also connected the probe-gun anode (An′), one of the deflector plates (Df) and the fluorescent target (Ta). Behind the target, there are indicated in the figure the two co-ordinate systems (y', z') of the probe gun and (x, z) of the space-charge producing gun.

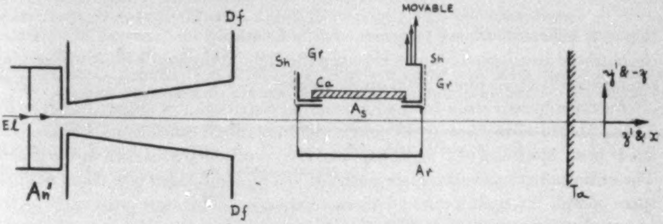
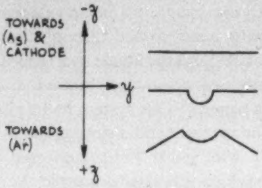


Figure 14. Probing of space charge.

Initially the cathode (Ca) was not heated, so that no space charge was produced in the anode box. As long as grid (Gr) and cathode (Ca) were kept at the potential (V_A), the cross-section of the probe beam could be seen on the target as a straight line, as shown in figure 15 (*a*). When the cathode (Ca) and grid (Gr) were charged up to -20 volts with respect to the anode voltage (V_A), the picture shown in figure 15 (*b*) could be seen on the fluorescent target. The bulge in the middle of the line was due to the bulging of the equipotentials in front of the anode slot (As), since the field between cathode and anode was penetrating through this slot. When, further, the cathode was heated by a proper filament current, its emission current was observed to produce the picture shown in figure 15 (*c*) on the target. The bulge had widened, and moved in the direction of the cathode, while the ends of the line kept approximately their original position. The movement of the bulge at the target was measured by a microscope within 0·1 mm. accuracy and found to be 3 mm. The movement in the anode box could be estimated to be less than 1 mm

The shift of the probe beam in the middle of the bulge is a significant indication for the field set up by the space charge along the z-axis of the anode box. The observed movement of it towards As showed that there must have been set up a potential minimum in the anode box, and that the probe beam was passing at the side of the minimum which gradually sloped towards the cathode.

Now, the whole space-charge producing gun was mounted on a slide, and by means of an external ground joint it could be screwed up and down in the z-direction as is indicated in figure 14 by arrows with the inscription " movable ". Thus, the potential distribution could be explored throughout the anode box by observing the movement of the probe beam on the target.



(*a*) No cathode potential difference.

(*b*) With cathode potential difference.

(*c*) With cathode heating,

Figure 15. Picture on fluorescent target.

An experimental result for the longitudinal distribution is shown in figure 16. There, plotted as abscissa, is the position of the probe beam on the z-axis of the anode box; anode slot (As) and anode roof (Ar) position are marked in the graph. The ordinate represents the space potential. The solid curve was taken without space charge, the broken curve with space charge. The close proximity of the potential minimum to the anode roof (Ar) seemed surprising, but can probably be explained by taking into account some secondary emission or electron reflection at Ar.

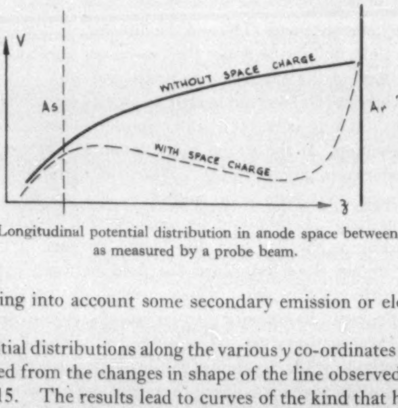


Figure 16. Longitudinal potential distribution in anode space between As and Ar, as measured by a probe beam.

Lateral potential distributions along the various y co-ordinates of the anode box could be estimated from the changes in shape of the line observed at the target, as shown in figure 15. The results lead to curves of the kind that have been shown already in figure 13. The potential minima obtained are deep enough to produce a strong inhomogeneity of electron velocities, The results of this section are only qualitative, but they clearly show two important phenomena, the building up of deep potential minima across and along an electron beam. Further experiments, however, will be needed to explain quantitatively the discrepancy between the experimentally observed current distribution and beam spread predicted by the simple theory.

ACKNOWLEDGMENTS

The work described in this paper was carried out on behalf of the Director of Scientific Research, Admiralty, and the author wishes to thank the Board of Admiralty for permission to publish. The author also wishes to acknowledge his indebtedness to Mr. I. Shoenberg, Director of Research, to Mr. G. E. Condliffe and Mr. L. F. Broadway, and to his colleagues in the Laboratories of Electric and Musical Industries, Ltd., Hayes, Middlesex, where the work was carried out.

REFERENCES

BOUWERS, A., 1935. *Physica*, **2**, 145.
BROADWAY, L. F. and BULL, C. S., 1940. British Patent No. 574,512.
BULL, C. S., 1945. *J. Instn. Elect. Engrs.*, **92**, 86.
BULL, C. S. and O. KLEMPERER, 1944. (Unpublished.)
KLEMPERER, O., 1939. *Electron Optics* (Cambridge : The University Press).
KLEMPERER, O., 1947. "Influence of space charge on thermionic emission velocities." *Proc. Roy. Soc.*, A. (In course of publication.)
KLEMPERER, O. and MAYO, B. J., 1947. "Electron optics and space charge in simple emission systems with circular symmetry." *J. Instn. Elect. Engrs.* (In course o publication.)
KLEMPERER, O. and WRIGHT, W. D., 1939. *Proc. Phys. Soc.*, **51**, 296.
LANGMUIR, I. and COMPTON, K. T., 1931. *Rev. Mod. Phys.*, **3**, 248.
PIERCE, J. R., 1939. British Patent No. 545,835.
PIERCE, J. R., 1940. *J. Appl. Phys.*, **11**, 548.
SEARS, J. B., 1933. *J. Sci. Instrum.*, **10**, 376.
SMITH, L. P. and HARTMAN, P. L., 1940. *J. Appl. Phys.*, **11**, 220.
STRONG, J., 1941. *Rev. Sci. Instrum.*, **12**, 213.
THOMPSON, B. J., 1943. *Proc. Inst. Radio Engrs.*, **31**, 485.
THOMPSON, B. J. and HEADRICK, L. B., 1940. *Proc. Inst. Radio Engrs.*, **28**, 319.
WATSON, E. E., 1927. *Phil. Mag.*, **3**, 849.
WEHNELT, A. and BLEY, H., 1926. *Z. Phys.*, **35**, 338.

A SERIES FOR THE STATIONARY VALUE OF A FUNCTION

BY T. SMITH, F.R.S.,
National Physical Laboratory, Teddington

MS. received 9 December 1946

ABSTRACT. A formula is given for the stationary value of a function of any number of variables in terms of the values of the function and its differential coefficients at any point in the neighbourhood of the stationary position.

THE following theorem is to be proved:—

Let f be a function of any number of independent variables x_1, x_2, x_3, ... Denote differentiation of any function with respect to these variables by the addition of the corresponding suffix, so that f_1, f_2, f_3, ... are the first differential coefficients and $f_{11}, f_{12}, f_{13}, \ldots f_{22}, f_{23}, \ldots$ the second differential coefficients of f. It is assumed that the Hessian of f does not vanish, so that the symmetrical matrix

$$\begin{bmatrix} f_{11} & f_{12} & f_{13} & \cdots \\ f_{21} & f_{22} & f_{23} & \cdots \\ f_{31} & f_{32} & f_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

has an inverse **H**. Let the operator **D** be defined by

$$-\mathbf{D}\phi = (f_1 f_2 f_3 \ldots)\mathbf{H} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \cdot \\ \cdot \end{bmatrix}$$

(where $\phi_1, \phi_2, \ldots$ are the differential coefficients of the operand ϕ with respect to $x_1, x_2, \ldots$) with the special convention that **D** does not operate on the first differential coefficients of f. Then if the series

$$f + \sum_1^\infty \frac{1}{(n+1)!}\mathbf{D}^n f = F$$

is convergent, the first differential coefficients of F with respect to all the variables vanish, and F represents a stationary value of f.

It will be noted that the values of the variables at the stationary position are not required.

Since every term of F but the first is of the second or a higher order in the first differential coefficients of f, F represents a stationary value of f when these coefficients are zero. When they are not zero the changes in the variables required to approach nearer to the stationary position are, to a first approximation, proportional to these coefficients, so that an expansion in powers of first differential coefficients is appropriate. Normally there is no difficulty in choosing values of the variables which make the first differential coefficients (but not the Hessian) small; convergence is then rapid. The accuracy of the formula may be established by showing that the contribution of any term to a first differential coefficient of F can be written as the sum of two terms, and that the contributions of successive terms cancel one another.

Using literal suffixes to denote differentiation with respect to typical variables, the formula may be written

$$F = f - \frac{1}{2!}f_a f_b f^{ab} + \frac{1}{3!}f_a f_b f_c f^{cp} f^{ab}{}_p - \frac{1}{4!}f_a f_b f_c f_d f^{dq}(f^{cp} f^{ab}{}_p)_q + \cdots,$$

where the convention is adopted that the appearance of a letter in a product both as a subscript and a superscript implies a summation for all variables. Superscribed letters identify elements of the matrix

$$\mathbf{H} = \begin{bmatrix} f^{11} & f^{12} & f^{13} & \cdots \\ f^{21} & f^{22} & f^{23} & \cdots \\ f^{31} & f^{32} & f^{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

which is necessarily symmetrical as it is the inverse of the Hessian matrix. From the definition of **H**

$$\left.\begin{aligned} f_{ab}f^{bc} &= 0 \quad (a \neq c), \\ &= 1 \quad (a = c). \end{aligned}\right\} \qquad \ldots\ldots(1)$$

Differentiating with respect to a variable g,

$$f_{abg}f^{bc} + f_{ab}f^{bc}{}_{g} = 0.$$

Since subscripts denote differentiations, the first term is unaltered if a and g are interchanged, and therefore the same change can be made in the second term, or

$$f_{bg}f^{bc}{}_{a} = f_{ak}f^{kc}{}_{g},$$

where for convenience different letters are used to denote the dummy variables. Multiply both sides of this equation by $f^{am}f^{n}{}_{g}$. Then

$$(f_{bg}f^{gn})f^{am}f^{bc}{}_{a} = (f_{ak}f^{am})f^{gn}f^{kc}{}_{g}.$$

Since the number of variables is finite we may sum the terms in whatever order we like. By (1) the bracketed terms are zero except when on the left $b = n$ and on the right $k = m$. With these values both bracketed sums are equal to unity, and therefore

$$f^{am}f^{cn}{}_{a} = f^{gn}f^{cm}{}_{g} = f^{an}f^{cm}{}_{a}, \qquad \ldots\ldots(2)$$

or the interchange of the indices m and n does not alter the value of the product.

This result may be generalized. Let X be any function of the elements of **H** such as occurs in the expression for F, and consider the product

$$f^{am}(f^{bn}X_{n})_{m},$$

which is equal to

$$f^{am}f^{bn}{}_{m}X_{n} + f^{am}f^{bn}X_{mn}.$$

By (2) the a and b in the first term may be interchanged, and, since X_{mn} is unaltered by interchanging m and n, the interchange of a and b in the second term also makes no change, or

$$f^{am}(f^{bn}X_{n})_{m} = f^{bm}(f^{an}X_{n})_{m}. \qquad \ldots\ldots(3)$$

Now apart from a numerical factor, each term in F is the product of two parts; the first contains only f's with a single subscript and no superscript; and the second f's with two superscripts, some of which are also differentiated. By repeated application of (3) we may make interchanges of the letters $a, b, c, \ldots$ (which do not imply summation when this second part is considered in isolation from the first) without altering the value of the part. When F is differentiated with respect to the variable h, the differentiation of the first part of any term is represented by substituting a double subscript $ah, bh, ch, \ldots$ for one of the single subscripts $a, b, c, \ldots$ Let the second part of the product be arranged with the corresponding factors f^{at}, f^{bt}, f^{ct}, $\ldots$ respectively in the leading position. Summation with respect to the common letter then leads by (1) to zero values except when $t = h$, and, whichever of the factors $f_{a}, f_{b}, f_{c}, \ldots$ is differentiated, the contributions to F_{h} are all equal. The total contribution from a term with $r+1$ factors in the first part is therefore $r+1$ times that resulting from the differentiation of only one of these factors, and this multiplier converts the numerical factor $\dfrac{1}{(r+1)!}$ to $\dfrac{1}{r!}$. The differentiation of the second part of any term is of course represented by the addition to it of the suffix *h*. Thus

$$\begin{aligned}F_h = f_h &- \frac{1}{2!}[2f_a f_{bh} f^{ab} + f_a f_b {f^{ab}}_h]\\ &+ \frac{1}{3!}[3f_a f_b f_{ch} f^{cp} {f^{ab}}_p + f_a f_b f_c (f^{cp} {f^{ab}}_p)_h]\\ &- \frac{1}{4!}[4f_a f_b f_c f_{dh} f^{dq}(f^{cp} {f^{ab}}_p)_q + f_a f_b f_c f_d \{f^{dq}(f^{cp} {f^{ab}}_p)_q\}_h]\\ &+ \dots\\ = f_h &- \frac{1}{2!}[2f_h + f_a f_b {f^{ab}}_h]\\ &+ \frac{1}{3!}[3f_a f_b {f^{ab}}_h + f_a f_b f_c (f^{cp} {f^{ab}}_p)_h]\\ &- \frac{1}{4!}[4f_a f_b f_c (f^{cp} {f^{ab}}_p)_h + f_a f_b f_c f_d \{f^{dq}(f^{cp} {f^{ab}}_p)_q\}_h]\\ &+ \dots\\ = 0&,\end{aligned}$$

the result to be proved.

The work described above has been carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory.

REVIEWS OF BOOKS

Researches on Normal and Defective Colour Vision, by W. D. WRIGHT. Pp. xvi+383. (London: Henry Kimpton, 1946.) 36*s*.

In this book are collected together the results of twenty years of research on colour vision by Dr. W. D. Wright and his co-workers at the Imperial College. The record includes several of the cardinal modern investigations on the subject: the determination of the colour coefficients of the spectrum colours for trichromats, anomalous trichromats and dichromats, measurements of the hue and saturation limens for all these classes and of the general colour limen for trichromats, the development of the method of binocular colour-matching for the study of colour adaptation and for the determination of the fundamental response curves of the trichromatic mechanisms, the comparison of luminosity curves in the fovea and parafovea at different brightness levels, and the quantitative study of colour and brightness matching in very small matching fields. Most of the measurements were carried out with the Wright trichromatic colorimeter, suitably modified where necessary, and an adequate account of this instrument and of the experimental technique is included. Two preliminary chapters summarize the knowledge of the visual process necessary for the proper understanding of what follows. These preliminary chapters apart, the investigation of workers outside the Imperial College group are considered only as far as is necessary for the discussion of the results of the author and his collaborators.

In re-presenting this work, together with some hitherto unpublished extensions of it, Dr. Wright has had the opportunity of pruning away the less significant material, of correcting a few erroneous conclusions and of reviewing the whole in the light of the latest information. All this is admirably done. The result is a clear exposition both of the basic experimental studies themselves and of the considered views of the author on their interpretation. As such, the book is indispensable to every research worker on vision.

It can also be strongly recommended to those concerned with the technological or medical aspects of colour.

Dr. Wright asks that his book should be regarded primarily as a record of experimental data. Much earlier work on colour vision—and even some current work—is of little value because of the inadequate specification of experimental conditions and results. It is very pleasant, therefore, to find a writer on the subject who foresees possible questions about the precise conditions of observation and takes care to provide the answers. One might wish perhaps that some of the curves showing the recovery of the eye after colour adaptation could have been plotted in terms of the fundamental stimuli (say Pitt's fundamental stimuli) instead of the instrumental primaries. Presumably this would not have altered any conclusions, but the discussion of the results would have been facilitated.

There are a few points in the exposition which are obscure or open to question. Insufficient emphasis is laid on the fact that heterochromatic brightness-matching (direct comparison) is an operation of a lower order of certainty than trichromatic colour matching. It is a pity that the notion of luminosity is not completely excluded from the original statement of the laws of trichromatic matching. Thus on page 108 the equation $x_1=u_1+v_1+w_1$ and the two similar equations are true only to the extent that additivity holds good in heterochromatic brightness-matching, and they might with advantage have been omitted in explaining a fundamental experiment on colour-matching. The use of " sensation magnitude " in the analysis of discrimination data (Chap. XVI) may be meat to some but will certainly be poison to others. While Dr. Wright is probably correct in assuming that matches made by the binocular method at various times after the removal of an adapting stimulus can be extrapolated back to give some property possessed by the retina while the adapting stimulus is still on, a rather fuller discussion of this point would have been welcome. A precise definition of the term " photo-chemical sensitivity " would be interesting. A statement on p. 345 suggests that because there is little change of hue with change of wave-length in a certain spectral region, radiations in that region will approximate to a fundamental stimulus acting on one trichromatic mechanism only. Surely there is the alternative that the spectral sensitivities of the mechanisms are in a constant ratio in the region in question, and this ratio may have any value. There seem to be very few minor slips, but one or two may be noted. Page 16, position of cone maximum, text and figure disagree ; p. 29, " film colour ", not " volume colour ", is the name usually applied (e.g. by Katz) to the illuminated field of an instrument ; p. 228, figure 137, caption in error ; p. 32, Hecht used two methods for arriving at the number of quanta absorbed by the visual purple at the threshold ; in the one which seems to be referred to here, the number was calculated from the sensitivity of the eye as a whole, not vice versa.

The book is well produced, but some variations in the paper provide a good example of the technological importance of small colour differences. W. S. S.

Introduction to Electron-Optics, by V. E. Cosslett. Pp. 272. (Oxford University Press, 1946.) 20*s*.

This is the first book to appear on this subject since the War. It is a welcome addition to the texts available to English workers, since it was written in the light of experience gained in imparting the principles of the subject to final-year university students. Such a task involves acquaintance with a wide field of work bearing on applications, some of which have been made throughout the war period, and are still in process of development.

The text is divided into two clearly defined and complementary sections—the theoretical and practical, and bears evidence of the wise discrimination exercised by the author in his choice of subject material. He has followed an established practice in presenting the properties of the electrostatic field. The methods used for solving some types of problem, including the relaxation method of Southwell, form an introduction to field plotting and ray tracing of trajectories. These, in turn, serve to introduce the focusing properties of all types of electrostatic lenses—aperture, cylinder, immersion and symmetrical—from the point of view of "geometrical optics ", which is a satisfactory approach. The chapter on magnetic focusing is especially well done, for here the author has contributed to some of the researches which he describes. The complicated subject of lens aberrations concludes the first section. These are treated individually, and the accompanying physical defects are clearly stated, but, for the full appreciation of the text, a " background " seems necessary here.

The second portion is mainly descriptive. It deals briefly with the fundamental requirements of electron devices, such as electron gun, image tube, multipliers, C.R. tube, electron microscope, diffraction, β-ray spectroscopy, magnetron, cyclotron, betatron, and velocity modulated tubes. Such an interesting review of development indicates the vast field of present-day electron-optical research ; it also makes both portions of the book blend well together. The general impression gained is that Dr. Cosslett has produced a well-balanced account of the principles and their applications.

Some parts could be expanded with profit, for example the effects of space charge on focused beams, electron distribution in beams (hardly touched on), design of practical guns arising from the latest work on " crossover " properties and the errors of deflecting fields. No doubt these will be dealt with in the next edition.

It is, nevertheless, an excellent monograph on the subject, and a distinct contribution to the literature.

L. JACOB.

Antennae: An Introduction to their Theory, by Dr. J. AHARONI. Pp. 265. (Oxford University Press, 1946.) 25*s*.

Although aerials have formed an essential and widely used element in radio communication from its very beginning, their theory is not sufficiently known even today. This is probably the reason why, until recently, there were practically no books dealing with the theory of aerials. During the last year, however, there has been a tendency to fill this gap. After the books of Pidduck and King, the book of Dr J. Aharoni means a serious effort to present as far as possible a complete survey of our theoretical knowledge on aerials.

The book begins with a clear exposition of electromagnetic theory by means of Maxwell's equations in a form appropriate for dealing with aerial circuits, using Hertzian vector and retarded electromagnetic potentials. The chapter is illustrated by a study of forced oscillations of a sphere and a prolate spheroid. The next chapter contains the theory of magnetic and electric dipoles, Hallèn's and King's solutions of current distribution and impedances of these aerials, some considerations on mutual impedance of aerials, the theory of receiving aerials, and polar diagrams of single aerials and aerial arrays. An interesting section deals with the effect of the earth, summarizing the work of Sommerfeld, Norton, Burrows, McPetrie and others. The last chapter presents Schelkunoff's theory of a dipole based on the biconical model.

The treatment of the subject in the book is purely mathematical, and it will not be easy reading for a radio engineer.

The author's intention to give an impartial survey of existing theories " without reflecting any opinion on the relative values of different methods " is a merit of the book on the one hand, but its weakness on the other, because it leaves the reader with an impression that our knowledge of aerials is a very complete one—which is far from the truth. In fact, most of the solutions for aerials of finite thickness are given in the form of infinite series, whose convergence has not been checked. J. H. Tait, of S.R.D.E., has shown, for example, in an as yet unpublished paper, that the classical Hallèn's solution of a dipole with an infinitesimal gap is divergent, and finite results for input impedance obtained by Hallèn, as well as by others using similar methods (King, Harrison, Blake, etc.), are probably due only to the approximations introduced into the original equation.

It is to be regretted that the author does not mention some recent experimental results, some of which do not quite confirm the validity of the existing theoretical results.

In spite of this criticism, the book is a valuable contribution to the literature of the subject.

B. STARNECKI.